# Analyzing Housing Data for Boston

**Authors: Om Mehta, Kirat Arora, John Wang,** *University of Massachusetts, Amherst, U.S.*

**Professor: Jonathan Larson,** *University of Massachusetts, Amherst, U.S.*

## 1 Introduction

Housing, like any other market, is driven by the economic principles of demand and supply. The price of housing is usually modeled by a hedonic model, as reviewed by Wing and Chin (2003); which identifies different factors and characteristics that affect the pricing of any item in the market. Our primary research question is to incorporate features from the Boston Housing Dataset and build a predictive model that can estimate the median price of a house.

In an urban environment, when crime rates go up, property prices go down Wei-Shong Lin (2014). Thaler (1978) showed that with a one-standard deviation increase in crime, the property prices in Rochester, NY went down by 3%. These studies support the claim that crime rates appear to be an important factor that quantifies the "safety" of a neighborhood. Crime, prima facie, appears to be a good measure of the demand for housing in any particular area and by extension, the price of the house. Svensson (2013) found that a permanent 1% increase in tax rates can cause the property price to fall permanently by 10%. Another parameter is the effect age plays on the value of a house. Coulson and McMillen (2008) break down the idea of house age into two variables: the age of the house, and the year it was built (which they term "vintage effects").

The Boston Housing Dataset serves as a crucial repository for unraveling the dynamics of the real estate market in the Boston, Massachusetts region. We aim to find a predictive model that can effectively help forecast housing prices based on the socio-economic factors in the diverse neighborhoods of Boston.

# 2 Methods

## 2.1 Dataset

We utilised data from the Boston Housing Dataset, originally derived from Harrison and Rubinfeld (1978). The authors drew the data from a U.S. Census Service Report concerning housing in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. This dataset has 506 records, where each data point represents a census tract in the Boston SMSA for owner-occupied one-family houses.

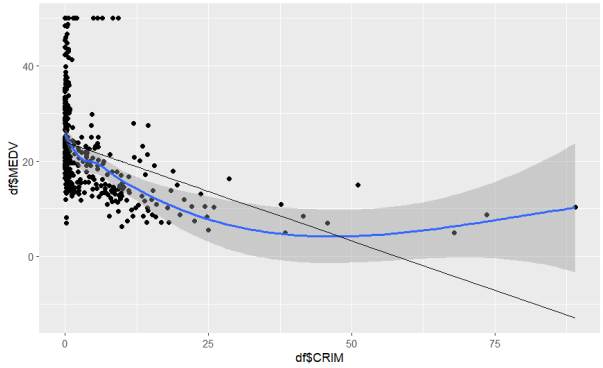
## 2.2 Variables

- **CRIM (Per Capita Crime Rate)**: The per capita crime rate by town.
- **ZN (Proportion of Residential Land)**: The proportion of residential land zoned for lots over 25,000 sq.ft.
- **INDUS (Non-Retail Business Proportion)**: The proportion of non-retail business acres per town.
- **CHAS (Charles River Dummy Variable)**: A binary variable (1 if the tract bounds the Charles River; 0 otherwise).
- **NOX (Nitric Oxides Concentration)**: The concentration of nitric oxides in parts per 10 million.
- **RM (Average Number of Rooms)**: The average number of rooms per dwelling.
- **AGE (Proportion of Older Units)**: The proportion of owner-occupied units built prior to 1940.
- **DIS (Weighted Distances to Employment Centers)**: The weighted distances to five Boston employment centers.
- **RAD (Accessibility to Radial Highways)**: An index of accessibility to radial highways.

- **TAX (Property Tax Rate)**: The full-value property-tax rate per $10,000.
- **PTRATIO (Pupil-Teacher Ratio)**: The pupil-teacher ratio by town.
- **B (Proportion of Blacks)**: A transformed variable based on the proportion of Black residents by town.
- **LSTAT (Percentage of Lower Status Population)**: The percentage of lower status population (not descrbibed in detail by the authors, but thought to be pertinent to the number of blue collar workers)
- **MEDV (Median Home Value)**: The median value of owner-occupied homes in $1000's.

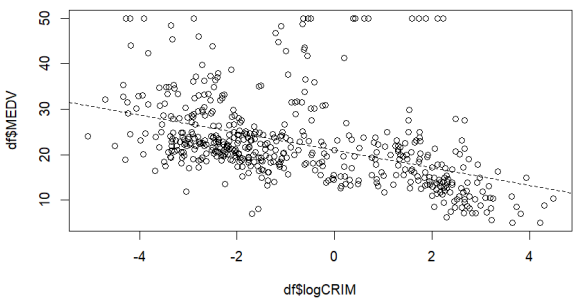
## 2.3 Transformation of variables

Figure 1: Transformation of crime




Plotting MEDV against CRIM gives us a clear non-linear plot - this is exarcebated by the presence of points with high CRIM values.

Figure 2: Transformation of crime



We clearly note a more linear relationship between the log-transformed CRIM and MEDV. The corresponding linear model also gives a higher $R^2$ when MEDV was regressed upon the transformed variable. For the sake of brevity, we state all similar transfomations that we performed - the code for which can be found in an accompanying file.

```
CRIM -> logCRIM

INDUS -> logINDUS

RAD.t -> 1 if (RAD == 24), 0 otherwise

DIS -> log(DIS)

LSTAT -> log(LSTAT)
```

## 2.4 Model

Input features: ZN, CHAS, RM, AGE, TAX, PTRATIO, logINDUS, RAD.t, logDIS, logLSTAT

Output feature: MEDV

# 3 Results

## 3.1 About the dataset

Figure 3: Histogram of the outcome variable

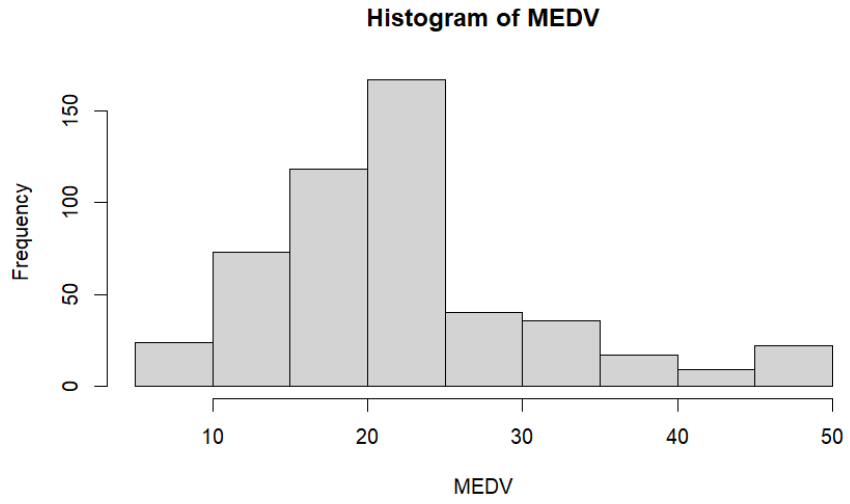


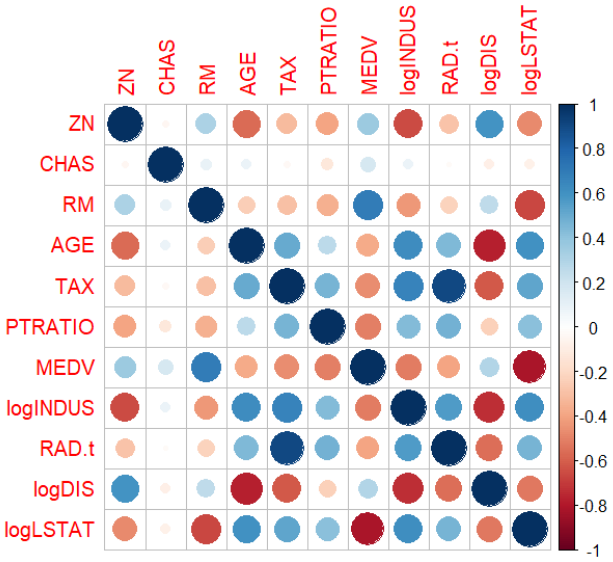

Figure 4: About our covariates

```
       ZN                RM              AGE              TAX        
 Min.   :  0.00   Min.   :3.561   Min.   :  2.90   Min.   :187.0  
 1st Qu.:  0.00   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.:279.0  
 Median :  0.00   Median :6.208   Median : 77.50   Median :330.0  
 Mean   : 11.36   Mean   :6.285   Mean   : 68.57   Mean   :408.2  
 3rd Qu.: 12.50   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.:666.0  
 Max.   :100.00   Max.   :8.780   Max.   :100.00   Max.   :711.0  
    PTRATIO          MEDV          logINDUS          logDIS      
 Min.   :12.60   Min.   : 5.00   Min.   :-0.7765   Min.   :0.1219  
 1st Qu.:17.40   1st Qu.:17.02   1st Qu.: 1.6467   1st Qu.:0.7420  
 Median :19.05   Median :21.20   Median : 2.2711   Median :1.1655  
 Mean   :18.46   Mean   :22.53   Mean   : 2.1602   Mean   :1.1880  
 3rd Qu.:20.20   3rd Qu.:25.00   3rd Qu.: 2.8959   3rd Qu.:1.6464  
 Max.   :22.00   Max.   :50.00   Max.   : 3.3229   Max.   :2.4954  
    logLSTAT    
 Min.   :0.5481  
 1st Qu.:1.9387  
 Median :2.4301  
 Mean   :2.3710  
 3rd Qu.:2.8306  
 Max.   :3.6368  
```

| | | |
|---|---|---|
| CHAS = 1 | 35 | 6.9% |
| CHAS = 0 | 471 | 93.1% |
| RAD.t = 1 | 132 | 26.1% |
| RAD.t = 0 | 374 | 73.9% |

Table 1: About our covariates: Categorical variables

Figure 5: Correlation matrix of our variables




## 3.2 Our model

Figure 6: Summary of our linear model

```
lm(formula = MEDV ~ . - CRIM - INDUS - DIS - LSTAT - RAD - NOX -
    logCRIM, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9648  -2.5931  -0.1625   1.9236  25.2523

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.175534   4.411135  11.601  < 2e-16 ***
ZN          -0.002740   0.012814  -0.214 0.830786
CHAS         2.644309   0.807433   3.275 0.001131 **
RM           2.675722   0.404891   6.609 1.01e-10 ***
AGE          0.002760   0.012666   0.218 0.827608
TAX         -0.011189   0.003272  -3.420 0.000679 ***
PTRATIO     -0.597610   0.117767  -5.075 5.51e-07 ***
logINDUS    -1.235433   0.491757  -2.512 0.012312 *
RAD.t        2.399569   1.131366   2.121 0.034423 *
logDIS      -4.254738   0.727624  -5.847 9.07e-09 ***
logLSTAT    -9.745038   0.589037 -16.544  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.438 on 495 degrees of freedom
Multiple R-squared:  0.7717,    Adjusted R-squared:  0.7671
F-statistic: 167.3 on 10 and 495 DF,  p-value: < 2.2e-16
```

## 3.3 About the residuals

Figure 7: Histogram of the residuals

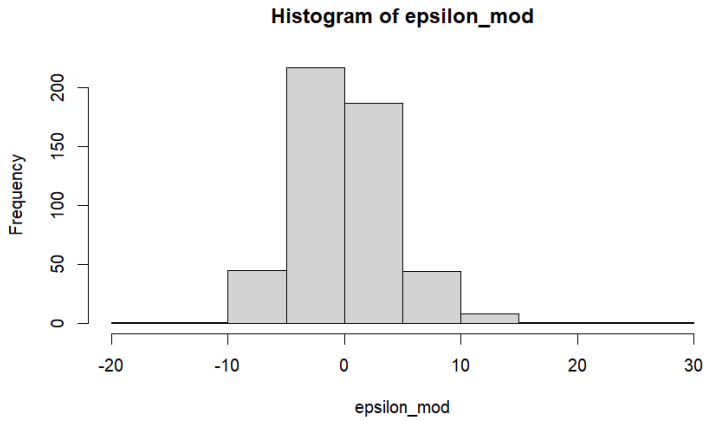



Figure 8: Boxplot of the residuals

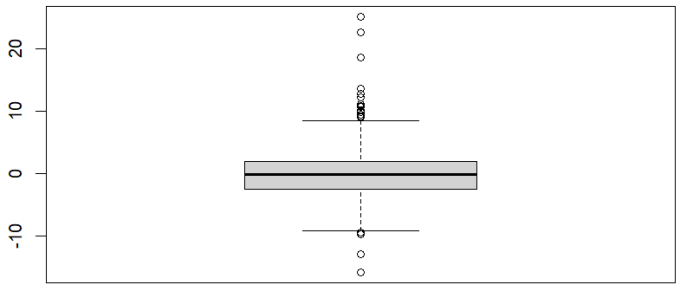



Figure 9: QQplot of the residuals

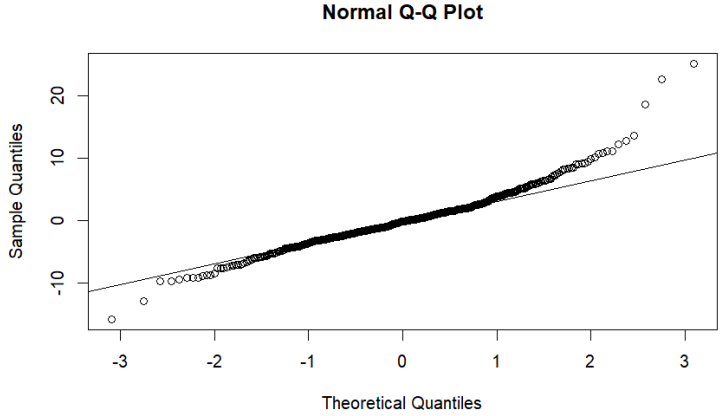
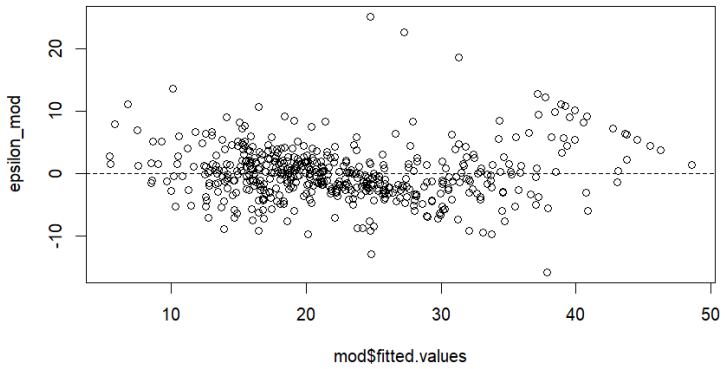


Figure 10: Residuals v/s fitted values



# 4 Discussion

## 4.1 How the study answers our research question

This study gives a pricing model that predicts our outcome (median housing price) with a reasonable accuracy. By means of the coefficients that we interpret below, we have also gained an idea of how each factor is related to the housing price - a key conclusion in any hedonic model.

## 4.2 Interpretation of some important coefficients

- If a tract surrounds the Charles river, a house will cost $2644 more on average than when it does not.
- An increase by one, of the median number of rooms in properties in a tract leads to the increase of the median price of house in that tract by $2675
- A 10% increase in "weighted distance from five Boston employment centres" in a tract decreases the median price in that tract by, on average, $405
- A 10% increase in the proportion of non-retail business acres per town decreases the median price in that tract by approximately $118 on average

## 4.3 Assumptions of linear regression

We noticed that our residuals were symmetric - had appromixately mean zero - and constant variance no matter the fitted values. The QQPlot, however, clearly shows that the residuals are not normally distributed. This means that any p-values should not be incorporated as probabilities. We can, however, interpret the coefficients as we did in terms of unit change per covariate.

## 4.4 Limitations of our study

Our models have an $R^2$ of around 75% - this corresponds to our model having explained around 75% of the variance displayed in this dataset. This means that there is still quite some variance that is unexplained by our linear model. The multiple transformations needed in order to achieve this relationship, combined with the fact that our residuals do not satisfy normality assumptions might be a sign of a linear model not being the best way to describe this data. Our model also utilises a lot of covariates which could be very specific to (and potentially be overfitting on) our training set.

## 4.5 Next Steps

We believe that the variance could potentially be explained better by using more flexible methods of curve fitting which could include machine learning techniques like random forests, decision trees, etc. Utilising a training-test split could also be looked at for model selection.

We also believe that alternative models hold great promise in further work in interpreting these statistics for real-world interpretations. Logistic regression, conducted on a more appropriate outcome, such as a transformed variable defined as 1 (safe) if CRIM<5; 0 (unsafe) otherwise - this could help in computing interesting statistics like the probability that we can secure a house that we can be 95%, based upon predictors like MEDV and DIS.

# References

Coulson, N. E. and McMillen, D. P. (2008). Estimating time, age, and vintage effects in housing prices. *Journal of Housing Economics*, 17:138–151.

Harrison, D. J. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage.; (United States)*, 5(1).

Svensson, L. E. (2013). The effect on housing prices of changes in mortgage rates and taxes. Technical report, Swedish House of Finance, Stockholm School of Economics.

Thaler, R. (1978). A note on the value of crime control: Evidence from the property market. *Journal of Urban Economics*, 5(1).

Wei-Shong Lin, Jen-Chun Tou, S.-Y. L. M.-Y. Y. (2014). Effects of socioeconomic factors on regional housing prices in the usa. *International Journal of Housing Markets and Analysis*, 7(1).

Wing, C. K. and Chin, T. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications*, 27:145–165.