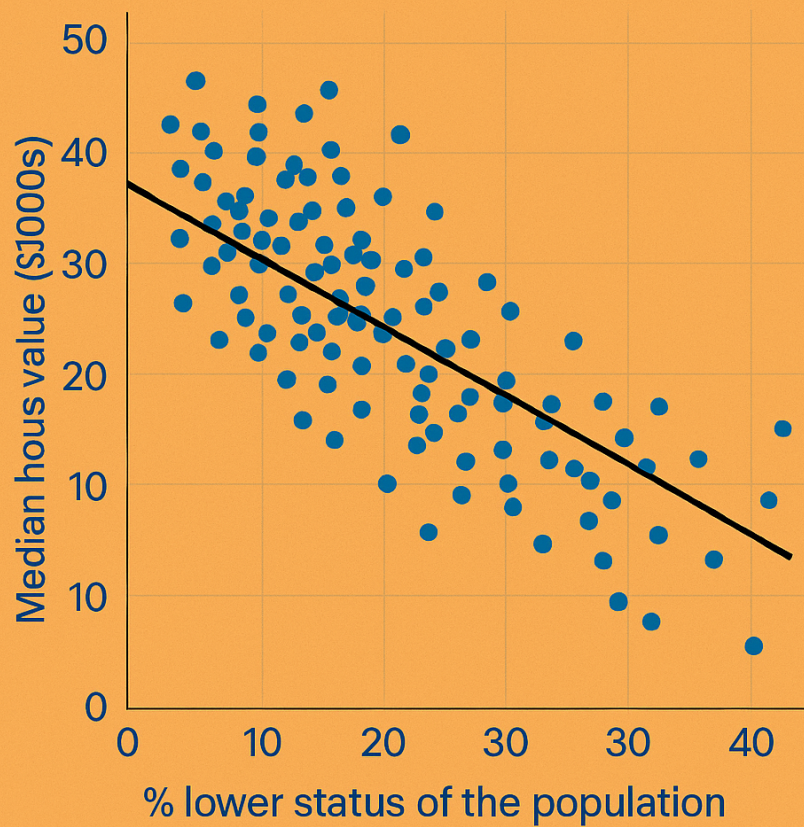


Predicting Median House Value in Boston



Predicting Median House Value in Boston

Seif H. Kungulio

November 01, 2025

Contents

1	Business Understanding	3
1.1	Introduction	3
1.2	Objectives	3
1.3	Problem Statement	3
1.4	Success Criteria	3
2	Data Understanding	4
2.1	Data Collection	4
2.2	Data Description	4
2.3	Data Dictionary	5
2.4	Initial Data Exploration	5
3	Data Preparation	6
3.1	Handle missing data	6
3.2	Transform skewed data	6
3.3	Encode categorical variables	6
3.4	Feature scaling	6
3.5	Split the data	6
4	Modeling	7
4.1	Multiple Linear Regression	7
4.2	Principal Component Regression (PCR)	7
4.3	XGBoost Regressor	7
5	Evaluation	8
6	Deployment	9
7	Conclusion	10
8	References	10

1 Business Understanding

1.1 Introduction

1.2 Objectives

The goal is to predict the median house value (MEDV) in Boston suburbs based on multiple explanatory variables such as crime rate, property tax, and number of rooms per dwelling.

1.3 Problem Statement

Problem statement

Boston's housing authority wants to understand the key factors influencing housing prices and build an accurate model to estimate property values for decision-making, urban planning, and investment.

1.4 Success Criteria

- Predictive: Achieve high accuracy (e.g., $R^2 > 0.80$) and low error (e.g., $RMSE < 3$) on unseen data.
- Business: Identify the top predictors of house price to guide policy or investment.

2 Data Understanding

2.1 Data Collection

The dataset used in this analysis is the Boston Housing Dataset, which contains information about various attributes of houses in Boston suburbs. The dataset is publicly available and can be accessed through the MASS package in R.

```
data("Boston", package = "MASS")
boston_data <- as_tibble(Boston)
glimpse(boston_data)
```

```
## Rows: 506
## Columns: 14
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.08829, ~
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5, 1~
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87, 7.~
## $ chas    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.524, ~
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.631, ~
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9, 9~
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9505~
## $ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, ~
## $ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311, 31~
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2, 15~
## $ black   <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, 394.12, 395.60, 396.90~
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17.10~
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9, 15~
```

2.2 Data Description

The dataset includes the following variables:

- MEDV: Median value of owner-occupied homes in \$1000s (target variable)
- CRIM: Per capita crime rate by town
- ZN: Proportion of residential land zoned for lots over 25,000 sqft
- INDUS: Proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX: Nitric oxides 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 centres
- RAD: Index of accessibility to radial highways
- TAX: Full-value property tax rate per \$10,000

- PTRATIO: Pupil-teacher ratio by town
- B: $1000(B_k - 0.63)^2$ where B_k is the proportion of
- LSTAT: Percentage of lower status of the population

2.3 Data Dictionary

Variable	Description.	Data Type	Constraints
MEDV	Median value of owner-occupied homes in \$1000s	Numeric	
CRIM	Per capita crime rate by town	Numeric	
ZN	Proportion of residential land zoned for lots over 25,000 sqft	Numeric	
INDUS	Proportion of non-retail business acres per town	Numeric	
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	Binary	
NOX	Nitric oxides concentration (parts per 10 million)	Numeric	
RM	Average number of rooms per dwelling	Numeric	
AGE	Proportion of owner-occupied units built prior to 1940	Numeric	
DIS	Weighted distances to five Boston employment centres	Numeric	
RAD	Index of accessibility to radial highways	Numeric	
TAX	Full-value property tax rate per \$10,000	Numeric	
PTRATIO	Pupil-teacher ratio by town	Numeric	
B	$1000(B_k - 0.63)^2$ where B_k is the proportion of black residents by town	Numeric	
LSTAT	Percentage of lower status of the population	Numeric	

2.4 Initial Data Exploration

3 Data Preparation

3.1 Handle missing data

3.2 Transform skewed data

3.3 Encode categorical variables

3.4 Feature scaling

3.5 Split the data

4 Modeling

4.1 Multiple Linear Regression

4.2 Principal Component Regression (PCR)

4.3 XGBoost Regressor

5 Evaluation

6 Deployment

7 Conclusion

8 References