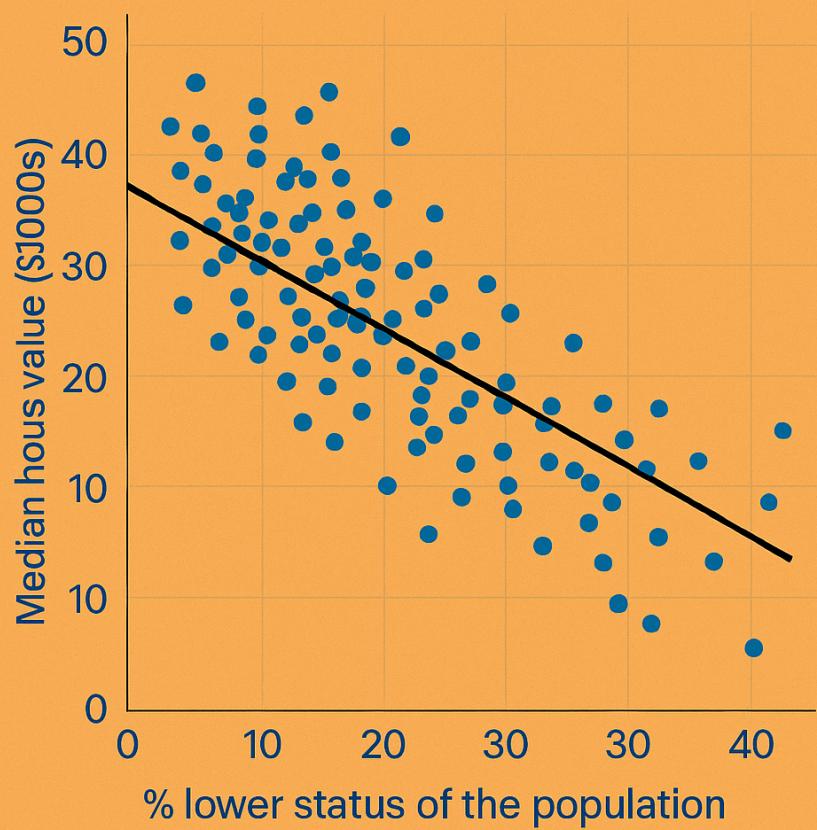


Predicting Median House Value in Boston



Predicting Median House Value in Boston

Seif H. Kungulio

November 03, 2025

Contents

1 Business Understanding	3
1.1 Introduction	3
1.2 Objectives	3
1.3 Problem Statement	3
1.4 Success Criteria	4
2 Data Understanding	5
2.1 Data Collection	5
2.2 Data Description	5
2.3 Data Dictionary	5
2.4 Key Observations	6
2.5 Initial Data Exploration	6
3 Data Preparation	10
3.1 Handle missing data	10
3.2 Transform skewed data	10
3.3 Encode categorical variables	10
3.4 Feature scaling	10
3.5 Split the data	10
4 Modeling	11
4.1 Multiple Linear Regression	11
4.2 Principal Component Regression (PCR)	11
4.3 XGBoost Regressor	11
5 Evaluation	12
6 Deployment	13
7 Conclusion	14
8 References	14

1 Business Understanding

1.1 Introduction

The Boston housing market represents a complex interplay of social, economic, and environmental factors that directly affect property values. Understanding these dynamics is essential for informed decision-making in real estate, urban planning, and public policy.

This project seeks to analyze and predict the median value of owner-occupied homes (MEDV) in the Boston area using the Boston Housing Dataset from the MASS package in R. The dataset includes 506 observations and 14 attributes describing housing and neighborhood characteristics, such as per capita crime rate, average number of rooms, proximity to employment centers, and levels of air pollution.

Through the application of advanced analytical techniques—Multiple Linear Regression, Principal Component Regression (PCR), and XGBoost Regressor—the study aims to build accurate predictive models and identify the most influential variables that explain housing price variability. Following the CRISP-DM framework, the process encompasses business understanding, data exploration, data preparation, modeling, evaluation, and deployment phases.

1.2 Objectives

The main objectives of this project are as follows:

- Primary Objective: To develop predictive models capable of estimating the median value of owner-occupied homes (MEDV) in Boston suburbs based on socioeconomic and environmental predictors.
- Secondary Objectives:
 - To identify key variables influencing housing prices in Boston.
 - To compare model performance between Multiple Linear Regression, Principal Component Regression, and XGBoost algorithms.
 - To provide actionable insights for urban planners, real estate investors, and housing authorities to guide data-driven decisions.
 - To demonstrate a reproducible, end-to-end analytical workflow following the CRISP-DM methodology.

1.3 Problem Statement

Boston's housing authority and urban developers face the challenge of accurately estimating property values across different neighborhoods. Property valuation depends on multiple interrelated factors—such as crime rate, accessibility to major roads, environmental quality, and local amenities—which can be difficult to model using traditional statistical techniques alone.

This project aims to answer the following key questions:

1. Which socioeconomic and environmental factors most strongly influence housing prices in the Boston area?
2. How can predictive modeling improve the estimation of housing values compared to conventional assessment methods?
3. Can machine learning models such as XGBoost enhance prediction accuracy and interpretability for policy and investment use?

1.4 Success Criteria

To determine the success of this project, both predictive and business metrics are defined:

- Predictive Success:
 - Achieve $R^2 \geq 0.80$ and $RMSE \leq 3.0$ on the test dataset.
 - Ensure generalization through cross-validation and avoid overfitting.
 - Compare performance across regression techniques and select the most robust model.
- Business Success:
 - Identify the top five predictors of housing value to support policy and investment strategies.
 - Deliver interpretable insights for urban planners and real estate analysts to optimize housing development and pricing strategies.
 - Provide a scalable and reproducible model pipeline for future data integration or policy forecasting.

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.df <- as_tibble(Boston) |>
  clean_names()

head(boston.df)

## # A tibble: 6 x 14
##       crim      zn indus    chas     nox      rm    age     dis      rad      tax ptratio black
##       <dbl>    <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.00632    18   2.31     0 0.538   6.58  65.2   4.09     1  296   15.3   397.
## 2 0.0273     0    7.07     0 0.469   6.42  78.9   4.97     2  242   17.8   397.
## 3 0.0273     0    7.07     0 0.469   7.18  61.1   4.97     2  242   17.8   393.
## 4 0.0324     0    2.18     0 0.458   7.00  45.8   6.06     3  222   18.7   395.
## 5 0.0690     0    2.18     0 0.458   7.15  54.2   6.06     3  222   18.7   397.
## 6 0.0298     0    2.18     0 0.458   6.43  58.7   6.06     3  222   18.7   394.
## # i 2 more variables: lstat <dbl>, medv <dbl>
```

2.2 Data Description

The dataset used in this project is the Boston Housing Dataset, originally published in the Harrison and Rubinfeld (1978) study and included in the MASS package in R. It contains 506 observations and 14 variables describing socioeconomic, demographic, environmental, and structural characteristics of housing in Boston suburbs. The dataset is commonly used for predictive modeling and regression analysis tasks, particularly for estimating housing values.

Each record represents a census tract in the Boston metropolitan area, with the target variable being the median value of owner-occupied homes (MEDV) expressed in \$1000s.

The features cover a wide range of factors that influence property values — including crime rates, industrial activity, environmental quality, accessibility to major roads, and education levels. These variables provide a multi-dimensional view of the housing environment, making the dataset suitable for developing regression and machine learning models to predict housing prices.

2.3 Data Dictionary

Variable	Description	Data Type	Constraints/Rule
crim	Per capita crime rate by town	Numeric	≥ 0
zn	Proportion of residential land zoned for lots over 25,000 sq.ft.	Numeric	$0 \leq \text{zn} \leq 100$
indus	Proportion of non-retail business acres per town	Numeric	≥ 0

Variable	Description	Data Type	Constraints/Rule
chas	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	Integer	0 or 1
nox	Nitric oxides concentration (parts per 10 million)	Numeric	$0 < \text{nox} \leq 1$
rm	Average number of rooms per dwelling	Numeric	> 0
age	Proportion of owner-occupied units built prior to 1940 (%)	Numeric	$0 \leq \text{age} \leq 100$
dis	Weighted distances to five Boston employment centers	Numeric	> 0
rad	Index of accessibility to radial highways	Integer	1–24
tax	Full-value property-tax rate per \$10,000	Numeric	> 0
ptratio	Pupil-teacher ratio by town	Numeric	> 0
black	$(1000(\text{Bk} - 0.63)^2)$, where Bk is the proportion of Black residents by town	Numeric	≥ 0
lstat	Percentage of lower status of the population	Numeric	$0 \leq \text{lstat} \leq 100$
medv	Median value of owner-occupied homes in \$1000s (Target variable)	Numeric	> 0 (typically capped at 50 in dataset)

2.4 Key Observations

- The dataset does not contain missing values, making it well-suited for direct modeling.
- Some predictors (e.g., CRIM, LSTAT, RM) exhibit skewness that may require transformation.
- The target variable (MEDV) is continuous and typically ranges between \$5,000 and \$50,000, with values above 50 capped in the dataset.
- There is potential multicollinearity between some predictors (e.g., RAD and TAX), which will be explored during modeling and addressed using dimensionality reduction or regularization techniques.

2.5 Initial Data Exploration

2.5.1 Basic coercions for exploration

```
boston.df <- boston.df |>
  mutate(chas = factor(chas, levels = c(0, 1),
                       labels = c("No", "Yes")),
         rad = as.integer(rad))

glimpse(boston.df)
```

```
## 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, 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.~
## $ chas    <fct> No, No~
## $ 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, 5, 4, 4, 4, 4, 4, 4, 4, 4, ~
## $ tax      <dbl> 296, 242, 242, 222, 222, 222, 311, 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~
```

2.5.2 High-level summary

```
skim(boston.df)
```

Table 2: Data summary

Name	boston.df
Number of rows	506
Number of columns	14
<hr/>	
Column type frequency:	
factor	1
numeric	13
<hr/>	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
chas	0	1	FALSE	2	No: 471, Yes: 35

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
crim	0	1	3.61	8.60	0.01	0.08	0.26	3.68	88.98	
zn	0	1	11.36	23.32	0.00	0.00	0.00	12.50	100.00	
indus	0	1	11.14	6.86	0.46	5.19	9.69	18.10	27.74	
nox	0	1	0.55	0.12	0.38	0.45	0.54	0.62	0.87	
rm	0	1	6.28	0.70	3.56	5.89	6.21	6.62	8.78	
age	0	1	68.57	28.15	2.90	45.02	77.50	94.07	100.00	
dis	0	1	3.80	2.11	1.13	2.10	3.21	5.19	12.13	
rad	0	1	9.55	8.71	1.00	4.00	5.00	24.00	24.00	
tax	0	1	408.24	168.54	187.00	279.00	330.00	666.00	711.00	
ptratio	0	1	18.46	2.16	12.60	17.40	19.05	20.20	22.00	
black	0	1	356.67	91.29	0.32	375.38	391.44	396.22	396.90	
lstat	0	1	12.65	7.14	1.73	6.95	11.36	16.96	37.97	
medv	0	1	22.53	9.20	5.00	17.02	21.20	25.00	50.00	

```
summary(boston.df)
```

```

##      crim          zn         indus        chas        nox
##  Min. : 0.00632  Min. : 0.00  Min. : 0.46  No :471  Min. :0.3850
##  1st Qu.: 0.08205 1st Qu.: 0.00  1st Qu.: 5.19  Yes: 35  1st Qu.:0.4490
##  Median : 0.25651 Median : 0.00  Median : 9.69           Median :0.5380
##  Mean   : 3.61352 Mean  : 11.36  Mean  :11.14           Mean  :0.5547
##  3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10           3rd Qu.:0.6240
##  Max.   :88.97620 Max.  :100.00  Max.  :27.74           Max.  :0.8710
##      rm           age          dis          rad
##  Min. :3.561    Min. : 2.90    Min. : 1.130  Min. : 1.000
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  1st Qu.: 4.000
##  Median :6.208   Median : 77.50   Median : 3.207  Median : 5.000
##  Mean   :6.285   Mean  : 68.57   Mean  : 3.795  Mean  : 9.549
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  3rd Qu.:24.000
##  Max.   :8.780   Max.  :100.00   Max.  :12.127  Max.  :24.000
##      tax          ptratio        black        lstat
##  Min. :187.0    Min. :12.60    Min. : 0.32  Min. : 1.73
##  1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  1st Qu.: 6.95
##  Median :330.0   Median :19.05   Median :391.44  Median :11.36
##  Mean   :408.2   Mean  :18.46   Mean  :356.67  Mean  :12.65
##  3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  3rd Qu.:16.95
##  Max.   :711.0   Max.  :22.00   Max.  :396.90  Max.  :37.97
##      medv
##  Min. : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
##  3rd Qu.:25.00
##  Max.   :50.00

```

2.5.3 Univariate analysis

Skewness table

```

numeric_vars <- boston.df |> select_if(is.numeric) |> names()

skewness_values <- boston.df |>
  summarize(across(all_of(numeric_vars), ~ e1071::skewness(.))) |>
  pivot_longer(everything(), names_to = "Variable", values_to = "Skewness") |>
  arrange(desc(abs(Skewness)))

kable(skewness_values, digit=2, caption = "Skewness of numeric variables")

```

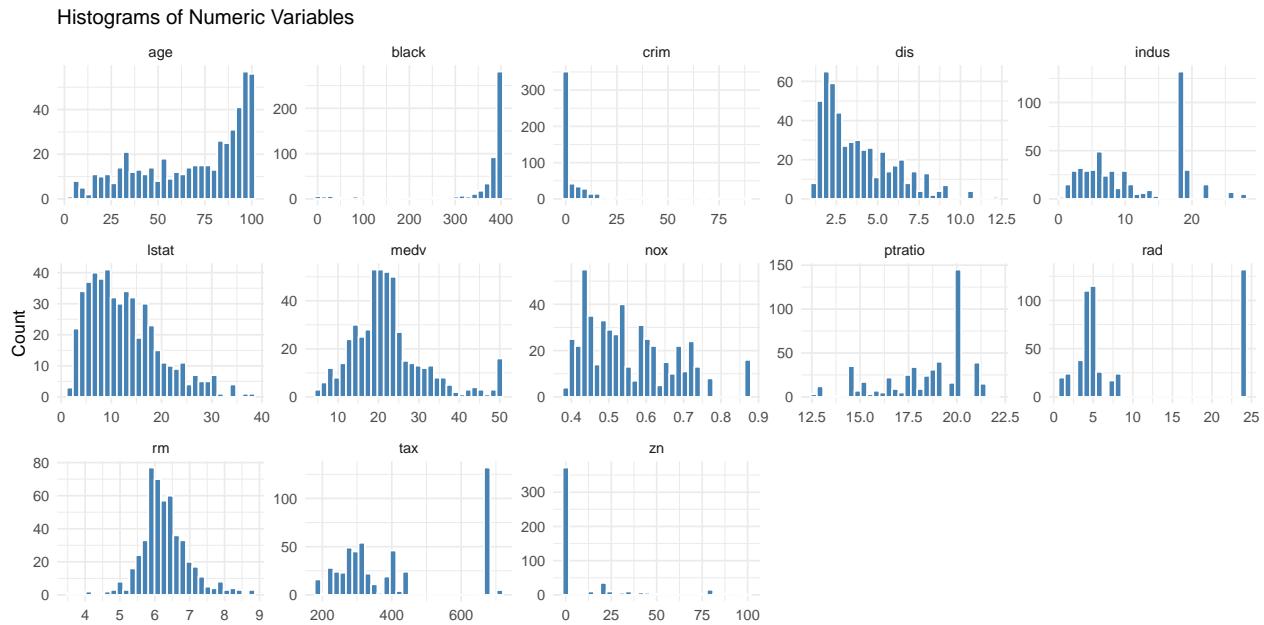
Table 5: Skewness of numeric variables

Variable	Skewness
crim	5.19
black	-2.87
zn	2.21
medv	1.10
dis	1.01
rad	1.00
lstat	0.90

Variable	Skewness
ptratio	-0.80
nox	0.72
tax	0.67
age	-0.60
rm	0.40
indus	0.29

Histograms of numeric variables

```
boston.df |>
  pivot_longer(cols = all_of(numeric_vars)) |>
  ggplot(aes(value)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  facet_wrap(~ name, scales = "free", ncol=5) +
  labs(title = "Histograms of Numeric Variables",
       x = NULL, y = "Count")
```



2.5.4 Bivariate analysis

Correlation matrix for numeric variables

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