# Predicting Median Value for Houses in Boston Census Tracts

Alex Jaimes Sandoval, Roshan Sanyal, and Megan Schaeb

December 09, 2022

**Executive Summary**

The housing market is responsible for a substantial portion of the U.S. gross domestic product and impacts millions of Americans who either own or are looking to own a home. In Boston, there is a large diversity of neighborhoods and a wide range of house prices. Given that different parts of the city experience different conditions, the question then arose of what factors in particular play a role in Boston house prices.

This report focuses on predicting median housing prices for census tracts in Boston, Massachusetts using data gathered during the 1970 U.S. Census, which is available on Kaggle. Although the data is not recent, and thus the predicted house prices are outdated, analysis of which environmental, social, and architectural factors impact house prices could give insight into homebuyers' values. Furthermore, repeating the analysis detailed in this report on a more recent dataset could help assess how homebuyers' values and priorities change over time.

Before fitting any models, a thorough exploratory data analysis was conducted to visualize relationships among median home value and the 13 other variables in the dataset. To check for multicollinearity and linear model assumption violations, a model was fit using all of the predictors, and the multicollinearity issue identified in the model was resolved by removing a variable. For the model selection process, k-fold cross-validation (with 10 folds) was used to split the 506 observations in the dataset into training and testing sets, with 90% of the data used for training and the remaining 10% for testing. Several linear regression models were then fit and compared under four criteria: Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), adjusted $R^2$, and out-of-sample mean squared error (MSE). The final model, which used the backward step selection process under BIC to remove variables, was chosen due to having the lowest AIC and lowest BIC of all the models evaluated. After selecting the final model, an investigation of the model assumptions showed a violation of normality among the residuals, which we have attributed to a lack of specific location data and the sensitivity of house prices to location.

The final model shows that a combination of several environmental, social, and architectural factors can be used to predict the price of a house in Boston. Further, the interaction terms in the model show that certain factors are related and jointly impact house prices. Going forward, an expert in the housing industry could help remove from the model any variables which represent contextually unjustified relationships and to uncover any hidden relationships which could have led to the inclusion of such unjustified variables in the model. Additionally, to gain an understanding of current house price indicators, the analysis should be replicated using more recent data.

## 1 Problem Context

As of 2022, the housing sector accounts for 16.4% of the total US gross domestic product (GDP) and contributes $2.8 trillion to the US economy (Dietz, 2022). With Boston, MA being one of the largest housing markets in the US, the question arose of what environmental, architectural, and social factors impact Boston house prices. To investigate this matter, data, which was collected during the 1970 US census and used in the 1978 journal article "Hedonic housing prices and the demand for clean air" (Harrison Jr. & Rubinfeld, 1978), was gathered from Kaggle (Fedesoriano, n.d.). This dataset contains 506 observations and 14 variables, including the median house value, for census tracts in Boston. Setting aside the median house value as the response variable, there are 13 potential explanatory variables in the dataset, 11 of which are quantitative and two that are categorical. Consequently, the data is sourced from the late 70s, which means the values are outdated. Moreover, the data does not take into account the actual locations of the homes. This is problematic because location is a critical factor in determining the actual asset value of the home.

## 2 Data Analysis

We started the analysis by exploring the data to understand which variables could serve as good predictors of the median price of Boston houses. Scatter plots and correlation coefficients computed between the quantitative predictors and the response revealed that some of them were strongly correlated, so we thought they would make good predictors. Similarly, boxplots and ANOVA tests were run on the categorical variables (see Table A-1). The analysis yielded the conclusion that the group means for the two categorical variables, index of radial highway access (RAD) and adjacency to the Charles River (CHAS), were different, indicating they likely have at least some predictive power. Given that we had 13 predictor variables and limited time, we did not look thoroughly at plots between the predictors.

Since the dataset contains so many variables and the response is continuous, we thought a multiple linear regression model would be ideal. The dataset contains a sizable number of observations, so we decided to set apart some data to validate the model and make sure overfitting was not an issue. We chose a 90-10 split of the data so that we would have roughly 50 observations to test the model.

We initially built a model using all of the data and containing all of the predictors in order to search for multicollinearity or any other glaring issues. This first model had multiple issues, such as heteroscedasticity and non-normality of the studentized residuals (see Figure A-4), and revealed a high number of outliers. After identifying the outliers, we attributed their status to simultaneous low weighted distances to Boston employment centers and high median house values (see Figure A-3). The initial model also showed that we had multicollinearity among

predictor variables. By looking at the variance inflation factor (VIF), found that the index of accessibility to radial highways (RAD) and full-value property-tax rate (TAX) both had a VIF over 10 (see Table A-2). We decided that we should remove the RAD variable since its VIF was close to 20, way higher than the usual threshold of 10. After the removal, we looked at the VIF of the variables again and concluded that the multicollinearity issue was gone since all VIF values were below five (see Table A-2).

After remedying the multicollinearity issue, we used 10-fold cross-validation to select a model. For each fold, we fit eight models, some of which included interaction terms, and some which were fit using the step selection process. We then evaluated these models using BIC, AIC, adjusted $R^2$, and out-of-sample mean squared error averaged over each fold (see Table A-3). Since the backward step selection method (using BIC to remove variables) with interaction terms had the lowest average AIC and BIC of the eight models, we chose this as the process by which to produce our final model.

The final model was trained using a random 90-10 split of all the data (the seed was set for replicability) and wound up having 12 individual predictors and 20 interaction terms. After fitting the model, we predicted the median house value using the remaining 10% of the data.

The predicted median house values were plotted against the actual values from the out-of-sample data to visualize the accuracy of the model. Figure 1 below shows a strong, linear relationship between the variables (correlation ≈ 0.95), indicating that the model has good predictive power.

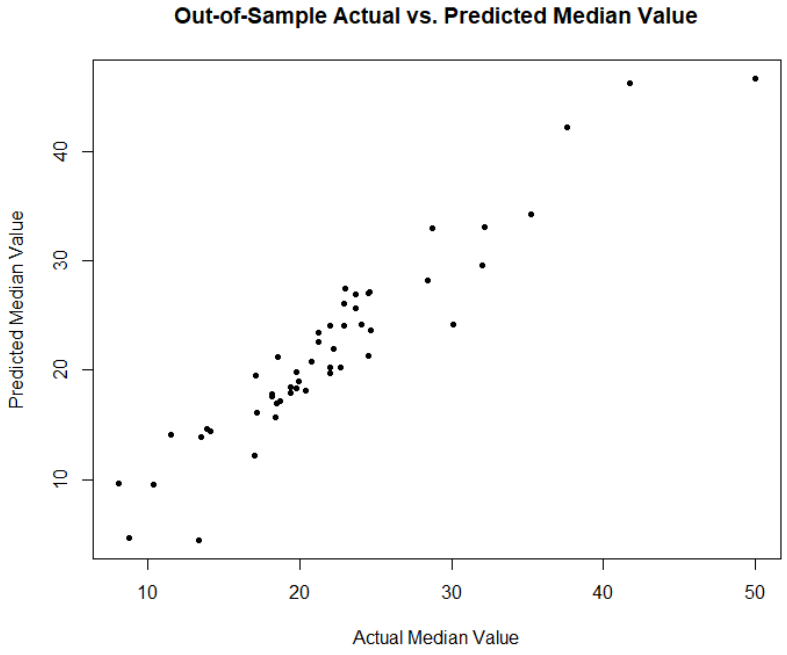


Figure 1: Actual vs. Predicted Values from Out-of-Sample Data

To determine the validity of our final model, we checked for violations of the homoscedasticity and normality assumptions. The Studentized Residuals vs. Fitted Values scatter plot in Figure 2 shows that the variance appears constant. This conclusion was confirmed using a Breusch-Pagan test, which produced a p-value of 0.05375. The Normal Q-Q Plot of the studentized residuals in Figure 2 and a p-value of 1.279e-7 from the Shapiro-Wilks test shows non-normality among the residuals.

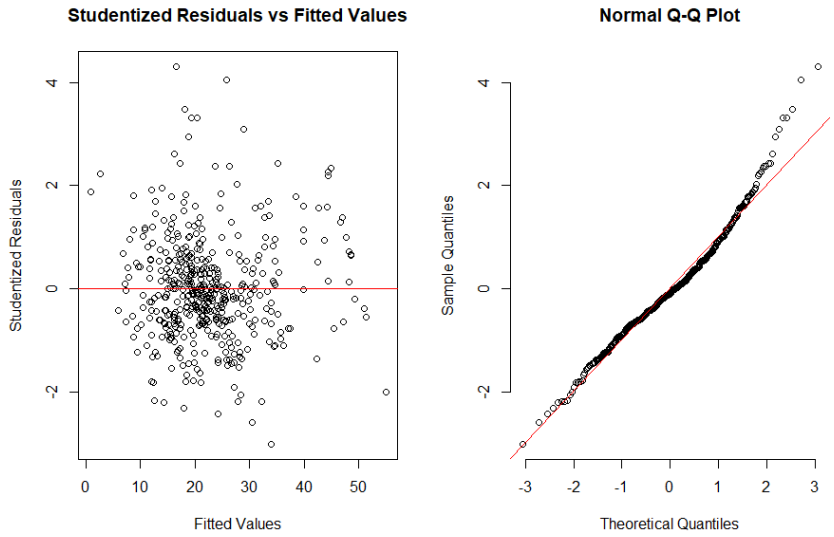


Figure 2: Residual Diagnostic Checking for Final Model

## 3 Conclusion

The final model ended up being able to predict the median house prices in Boston reasonably well. With its high adjusted R squared value, we can confidently say it can explain the majority of the variation in the data.

The main worry of the analysis is the lack of normality in the studentized residuals. Despite multiple attempts to address this issue, no real solution was found. We thought that a generalized linear model could be the solution to the normality problem, however, we were unable to estimate the true distribution of the response variable in order to really take advantage of the GLM's power. The lack of normality can be attributed to housing data being very erratic. This is because location is a large factor that can affect the value of homes and can subsequently lead to wild inconsistencies with the assets valuation. A Gaussian generalized linear model was tested, however, as was expected, the assumptions of that model were not met either due to the randomized quantile residuals not being normally distributed. This lack of normality in the residuals of the final model, of course, questions the validity of the analysis. For a future study, non-parametric techniques should be used to model the relationship between the response and its predictors.

## 4 Bibliography

Dietz, R. (2022, January 27). *Housing Share of GDP*. Eye on Housing. Retrieved December 9, 2022, from https://eyeonhousing.org/2022/01/housing-share-of-gdp-16-4/

Fedesoriano. (n.d.). *Boston House Prices-Advanced Regression Techniques*. Kaggle. Retrieved December 4, 2022, from https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data

Harrison Jr., D., & Rubinfeld, D. L. (1978, March). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81-102. https://doi.org/10.1016/0095-0696(78)90006-2

**Appendix**

The data included the following variables for Boston census tracts:
1) CRIM: per capita crime rate by town
2) ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3) INDUS: proportion of non-retail business acres per town
4) CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5) NOX: nitric oxides concentration (parts per 10 million) [parts/10M]
6) RM: average number of rooms per dwelling
7) AGE: proportion of owner-occupied units built prior to 1940
8) DIS: weighted distances to five Boston employment centers
9) RAD: index of accessibility to radial highways
10) TAX: full-value property-tax rate per $10,000 [$/10k]
11) PTRATIO: pupil-teacher ratio by town
12) B: The result of the equation B=1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13) LSTAT: % lower status of the population
14) MEDV: the median value of Boston houses

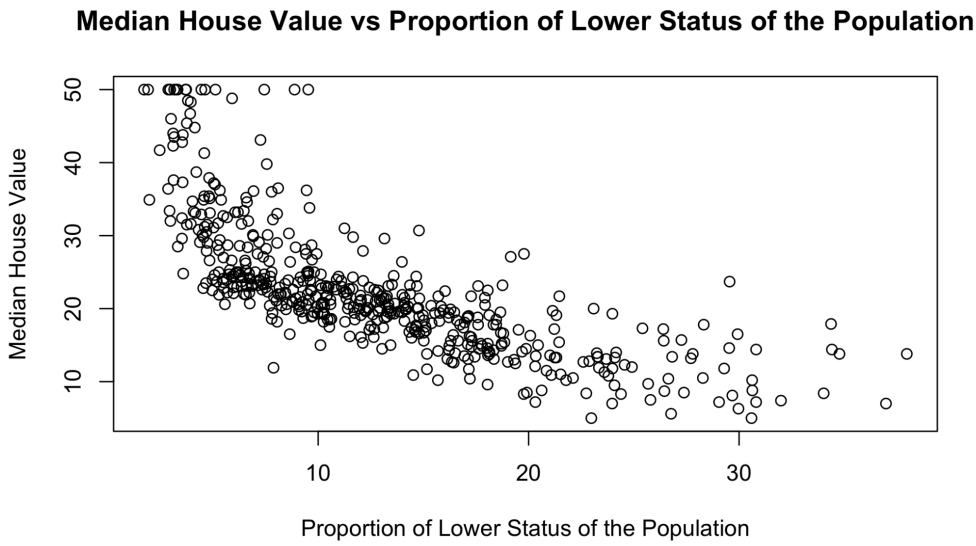


Figure A-1:Median House Value vs. Proportion of Lower Status of the Population (LSTAT)

Figure A-1 shows the relationship between the LSTAT variable and the response. The LSTAT variable represents the proportion of the population that is lower status = 1/2 (proportion of adults without some high school education and proportion of male workers classified as laborers). There is a clear trend in the data that could be used to contribute to the predictive model. This was one of numerous scatter plots produced to explore the data.

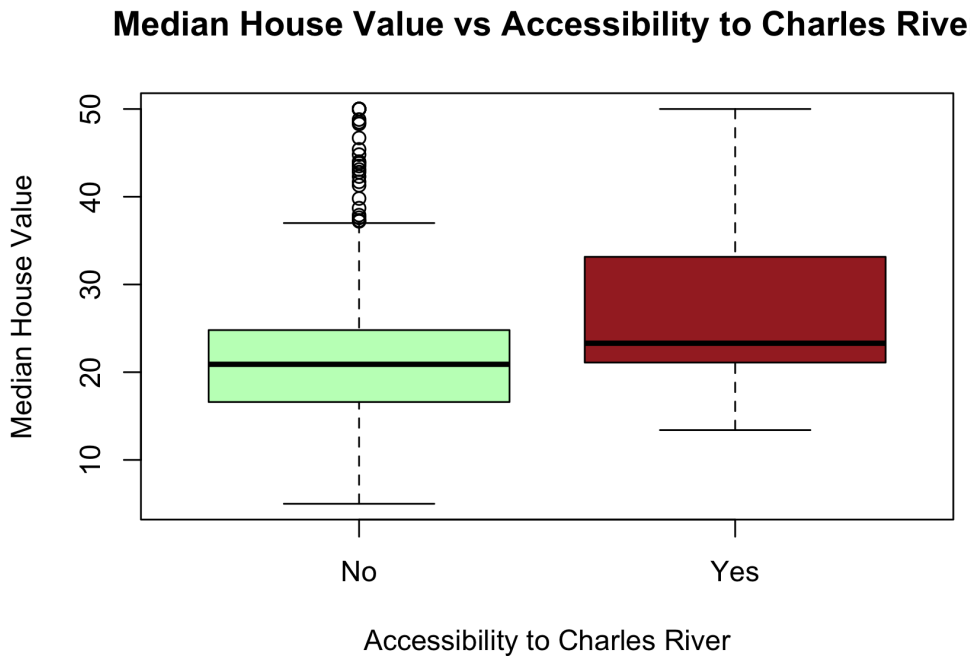


Figure A-2: Median House Value vs. Accessibility to Charles River

Figure A-2 above shows the distribution of median house value based on accessibility to the Charles River. From the plot, there does appear to be a difference in the distributions for the two groups, but we ran an analysis of variance (ANOVA) to determine if this difference was statistically significant. Table A-1 shows the results from this analysis, and the p-value of 7.39e-05 suggests there is a significant difference between the group means. As such, we thought that this categorical variable could be a useful predictor for house prices.

Table A-1: ANOVA on Accessibility to Charles River (CHAS)

| | Residual Degrees of Freedom | Sums of Squares | Mean Squares | F Statistic | P-Value |
|---|---|---|---|---|---|
| **CHAS** | 1 | 1312 | 1312.1 | 15.97 | 7.39e-05 |
| **Residuals** | 504 | 41404 | 82.2 | — | — |

After fitting a full model, two extreme outliers were identified and the source of the outlier behavior was attributed to a combination of low values in the weighted distance to Boston employment centers (DIS) coupled with high median house prices (MEDV). Figure A-3 shows these outliers on a plot of DIS versus MEDV.

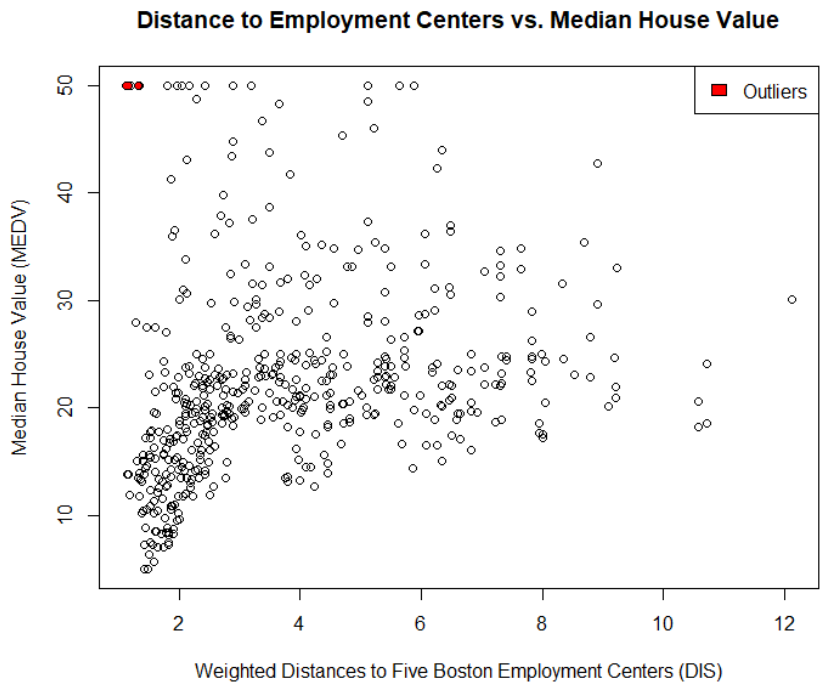


Figure A-3: Distance to Employment Centers vs. Median House Value

The Studentized Residuals vs. Fitted Values scatter plot and the Normal Q-Q Plot of the studentized residuals from the initial model, shown in Figure A-4, reveal nonconstant variance and non-normality, respectively. The constant variance assumption held up in our final model, however, the non-normality persisted.

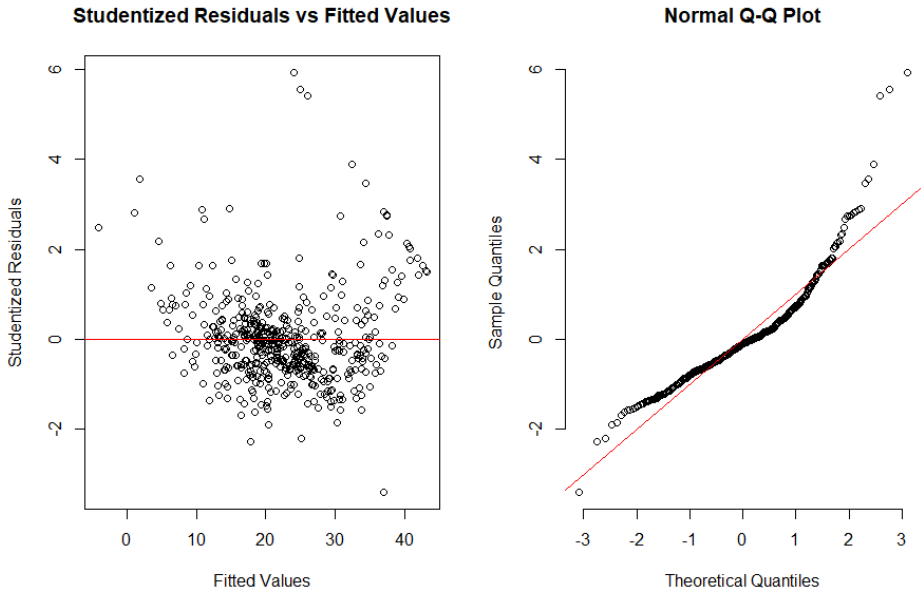


Figure A-4: Residual Diagnostic Checking for Initial Model

Calculating the VIF of the predictors in the initial model also revealed a relationship between RAD and TAX, as mentioned in Section 2. Both of these variables had a VIF over 10, but we chose to remove RAD from the model since the VIF was higher for that variable. As can be seen in Table A-2, excluding RAD from the model reduces the VIF values for each included predictor to an acceptable level (below 10).

| Initial Model | |
|---|---|
| | VIF |
| CRIM | 1.861919 |
| ZN | 2.405411 |
| INDUS | 4.542974 |
| CHAS | 1.099373 |
| NOX | 4.694575 |
| RM | 2.011739 |
| AGE | 3.158231 |
| DIS | 4.101432 |
| RAD | 20.049527 |
| TAX | 10.660453 |
| PTRATIO | 2.227791 |
| B | 1.326824 |
| LSTAT | 3.100999 |

| Without RAD | |
|---|---|
| | VIF |
| CRIM | 1.709995 |
| ZN | 2.211126 |
| INDUS | 3.658050 |
| CHAS | 1.046958 |
| NOX | 4.353862 |
| RM | 1.896570 |
| AGE | 3.067205 |
| DIS | 3.923741 |
| TAX | 3.584171 |
| PTRATIO | 1.730388 |
| B | 1.310980 |
| LSTAT | 3.031051 |

Table A-2: Variance Inflation Factor (VIF)

After removing RAD to deal with the multicollinearity in the initial model, 10-fold cross-validation was run on eight types of models. Table A-3 below shows the average AIC, BIC, adjusted $R^2$, and out-of-sample MSE for each of these models. For replicability, the seed was set to produce these results. As displayed in Table A-3, the backward step model with interactions (Model 6) had the lowest AIC and BIC, as well as the second lowest adjusted $R^2$ and out-of-sample MSE. The full model with interactions (Model 2) had the highest adjusted $R^2$ and out-of-sample MSE, as well as the second lowest AIC and BIC.

Since Model 6 contained only a subset of the predictors used in Model 2, an ANOVA F-test was conducted to determine if the additional predictors in Model 2 significantly improved the performance of Model 2 compared to Model 6. The results from that F-test, given in Table A-4 below, confirmed that the additional predictors in Model 2 were not significantly different from zero, and thus Model 6 (the simpler of the two models) was chosen as the final model.

Table A-3: K-Fold Cross Validation Results

| | AIC | BIC | Adjusted $R^2$ | Out-of-Sample Mean Squared Error |
|---|---|---|---|---|
| **[1] Full model** | 1448.653 | 1502.229 | 0.7303338 | 24.68688 |
| **[2] Full model with interactions** | 1063.662 | 1389.235 | **0.9132370** | **13.71792** |
| **[3] Backward step model** | 1446.602 | 1482.457 | 0.7264064 | 24.87432 |
| **[4] Forward step model** | 1446.602 | 1482.457 | 0.7264064 | 24.87432 |
| **[5] Mixed step model** | 1446.602 | 1482.457 | 0.7264064 | 24.87432 |
| **[6] Backward step model with interactions** | **1025.250** | **1175.673** | 0.9038579 | 14.37026 |
| **[7] Forward step model with interactions** | 1162.204 | 1238.033 | 0.8594015 | 15.40057 |
| **[8] Mixed step model with interactions** | 1162.302 | 1237.308 | 0.8592577 | 15.43850 |

Table A-4: ANOVA F-Test Model Comparison

| | Residual Degrees of Freedom | Sums of Squares | Degrees of Freedom | Mean Squares | F Statistic | P-value |
|---|---|---|---|---|---|---|
| **Model 6** | 420 | 3046.8 | — | — | — | — |
| **Model 2** | 374 | 2751.2 | 46 | 295.61 | 0.8736 | 0.7061 |