Final Report

Introduction

Understanding the factors that influence housing prices is critical for urban planning, policy development, and real estate management. The interplay among social, environmental, infrastructural and accessibility factors is complex, and these variables affect housing values in a dynamic manner. For instance, Ceccato and Wilhelmsson (2018) suggest that neighborhoods with high crime rates tend to experience lower housing prices. However, neighborhoods with high accessibility and located near the city centers may not see the same negative effects from crime. Additionally, Jones et al. (2023) observe that proximity to green spaces is positively associated with housing prices.

Although the impact of individual variables on housing prices has been widely studied, research on the combined effects of multiple factors remains deficient. This report aims to examine the relationship between various factors and the median value of owner-occupied homes in Boston neighborhoods.

Dataset description

The dataset used in this study was obtained from the StatLib archive (2016) and was originally published by Harrison and Rubinfeld (1978) to investigate the relationship between pollution and housing prices. The dataset was collected from diverse sources, including U.S. Census Service (1970), FBI (1970), and Massachusetts Department of Education (1971-1972) and others. The full list of sources is listed below along with the variable description.

The dataset has 495 observations, each representing a neighborhood in Boston. It includes 14 variables in total. The response variable in the regression analysis is the median value of owner-occupied homes (MEDV) in $1000's, collected from the 1970 U.S. Census. The remaining 13 variables are explanatory factors that capture structural, social, accessibility and environmental characteristics of each neighborhood.

1. CRIM - number of crimes per person in the town, data collected from the FBI
2. ZN - percentage of land in each town that is set aside for residential lots over 25,000 square feet, collected from the Metropolitan Area Planning Commission
3. INDUS - percentage of non-retail business land use, serves as an indication of consequences of industry: noise, heavy traffic, unpleasant view, collected from the Department of Commerce and Development
4. CHAS - dummy variable for if land touches Charles River (1 if yes; 0 otherwise), captures the amenity of a riverside location, collected by US Census

5. NOX - annual average of nitric oxide concentration in pphm (measured in parts per 10 million) collected from Department of Commerce and Development
6. RM - average number of rooms per home in a town, collected from US Census
7. AGE - proportion of owner-occupied units built prior to 1940 (older than 38 years in 1978), collected from US Census
8. DIS - average distance from five Boston employment centers (miles), collected by Harvard University
9. RAD - how easy to access highways from the neighborhood (standardized 1 to 24), collected from MIT
10. TAX - full-value property-tax rate per $10,000, measures the cost of public services in the neighborhood, collected from Massachusetts Taxpayers Foundation
11. PTRATIO - average number of students per teacher in each town, from Massachusetts Department of Education
12. B - measure related to proportion of Black residents in town: $1000(Bk - 0.63)^2$ where Bk is proportion, collected from US Census
13. LSTAT - proportion of lower income and adult without highschool education in each town, collected from US Census

Data Cleaning

The original dataset, boston.txt, contained observations split across two lines. To preprocess the data, we used the `readLines()` function to read the file and combined every two lines into one row. The dataset also included 22 rows of metadata and lacked headers; therefore, we skipped the first 22 rows while reading the data and then assigned appropriate column names to the dataset using `colnames(data)`. After loading the data, we checked for any missing values using the function `any(is.na(data))`. The check confirmed that there were no missing values (NA) in the dataset. With these steps completed, we obtained the cleaned data.

Analysis

Explorative Analysis

For the exploratory analysis, we first examined the summary statistics of the dataset. The descriptive statistics, including measures of central tendency, range, and standard deviation, did not reveal any unusual patterns or outliers. Overall, the variables appear to be well-behaved and exhibit generally expected characteristics. See *Appendix A* for more detailed summary statistics.

Next, we visualized the linear relationship between the full set of potential covariates and the response variable, MEDV (median housing prices), using scatterplots, as shown in *Figures 1.2* and *1.3*. Continuous covariates are displayed in scatterplots, while the categorical covariate, CHAS, is presented in a boxplot to assess the suitability of a linear model. From the plots, a linear relationship between several of our variables and the response variable, such as CRIM, NOX, RM, and LSTAT, is evident, along with some distinction between the two levels of CHAS. Looking at our correlation values (see *Table 1.1*), we find that the linear correlation is relatively moderate across the board, with absolute values generally ranging from 0.3 to 0.7. The lack of visual justification for including DIS (distance from employment centers), coupled with its low correlation value of 0.25, led us to exclude it during the exploratory phase.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
−0.3889686	0.3680704	−0.4820926	0.1785806	−0.4252230	0.6934430	−0.3743223	0.2490305	−0.3804263	−0.4670884	−0.5221533	0.3334051	−0.7386091	1.0000000

Table 1.1: Correlation between covariates and response

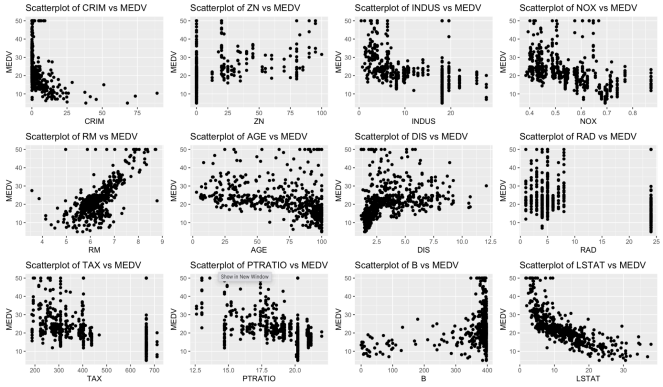


Figure 1.2: Scatterplot between covariates and response

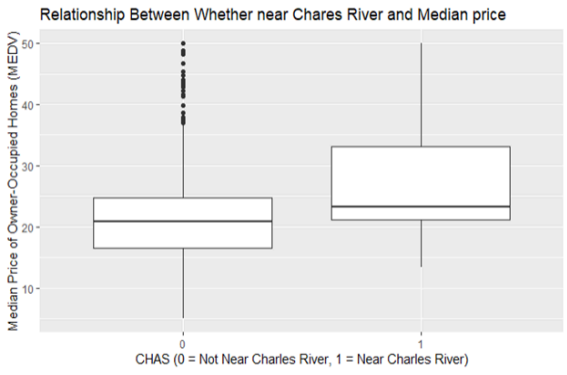


Figure 1.3: Boxplot of response based on CHAS

Finally, we assessed the multicollinearity issue by first examining pairwise scatterplots (see *Appendix B*) and identifying a couple of pairs with high correlations, such as the correlation between TAX and RAD (0.91) and the correlation between NOX and INDUS (0.763), which signals a potential issue of multicollinearity. Therefore, we further checked the Variance Inflation Factor (VIF) for the full set of potential covariates (see *Table 1.3*). However, none of these had a VIF greater than or equal to 10, indicating that multicollinearity is not a serious issue for the

model. As a result, we decided not to drop any covariates and proceeded with the current set of covariates.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	RAD	TAX	B	LSTAT
1.769837	1.944103	3.797346	1.075386	4.095525	1.892144	2.864293	7.437663	8.900614	1.815598	1.345440

Table 1.3: Variance Inflation Factor (VIF) values for each covariate

Statistical Analysis

Model selection

To construct a suitable model, we first performed both forward and backward selection techniques on our full set of potential covariates (see the two model selection steps in *Tables* 1.4 and 1.5). After completing both model selections, we found similarly high adjusted R^2 values of 0.69 for the forward-selected model and 0.70 for the backward-selected model, but a higher Mallows' Cp value for the forward-selected model (see *Table 1.6*). Given the similar performance of the forward-selected model with five covariates and the backward-selected model with ten covariates, we followed the principle of parsimony and chose the forward-selected model as our final model. The chosen model includes the covariates LSTAT (percentage of lower-income population), RM (rooms per home), PTRATIO (pupil-teacher ratio), B (percentage of Black residents), and CHAS (whether the property is near the Charles River).

Step	Variable dropped	*p*-value
1	ZN	0.648
2	INDUS	0.082

Table 1.4: Selection Steps for Backward Selection

Step	Variable added	*p*-value
1	LSTAT	0.0000
2	RM	0.0000
3	PTRATIO	0.0000
4	B	0.0002
5	CHAS	0.0003

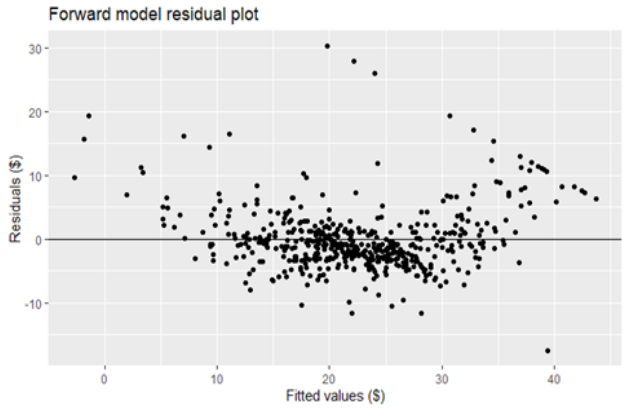
Table 1.5: Selection Steps for Foward Selection

Models	Mallow's Cp	Adjusted R^2
Forward-selected Model	26.12(p = 5)	0.6934
Backward-selected Model	12.24(p = 10)	0.7048

Table 1.6: Model Comparison in terms of Mallow's Cp and Adjusted R^2

Model Diagnostic

Then, we conducted Model diagnostic checks on the forward-selected model. First, the residual plot (see *Figure 1.6*) indicated a considerable disparity in the range of residuals. While most residuals are concentrated around -10 to 10, there are several extreme values that are significantly larger, which complicates the interpretation and separation of residuals. Second, the Q-Q plot (see *Figure 1.7*) highlighted a violation of the normality assumption, with evident skewness in the data. To address this issue, we decided to apply a log transformation to the response variable to restore normality and improve model fit assumption.



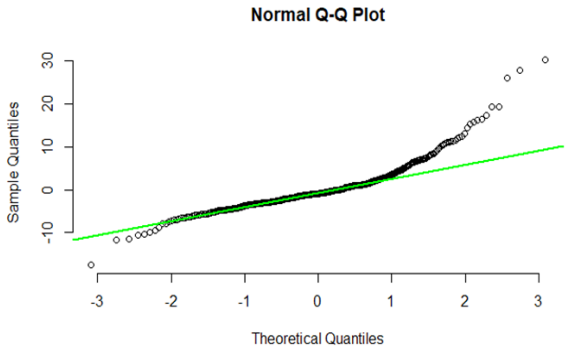


Figure 1.6: Q-Q plot for forward-selected model

Figures 1.7: Residual plots of forward-selected model

Data Transformation

After the log transformation, we observed that the residuals in the residual plot (see *Figure 1.8*) were more spread out and randomly centered around 0, as the scale had been adjusted by the log transformation. Our Q-Q plot (see *Figure 1.9*) also showed a more balanced tail, indicating the restoration of the normality assumption. However, the Q-Q plot still exhibited a heavier tail than a perfect normal Q-Q plot, suggesting that the error terms may follow a heavier-tailed distribution, such as a t-distribution or a Cauchy distribution. A secondary analysis using a serial residual plot (see *Appendix C*) confirmed that the independence assumption was not violated as well.

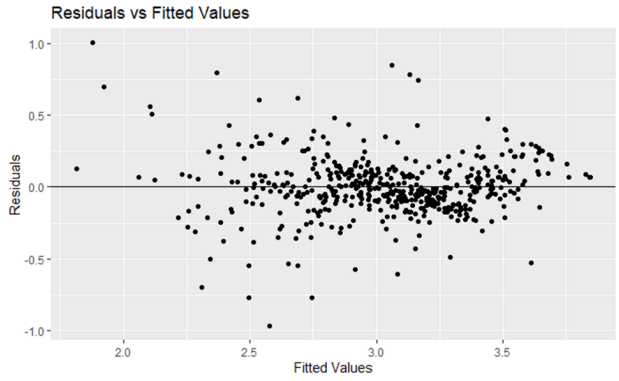


Figure 1.8: Residual plot of log-transformed model

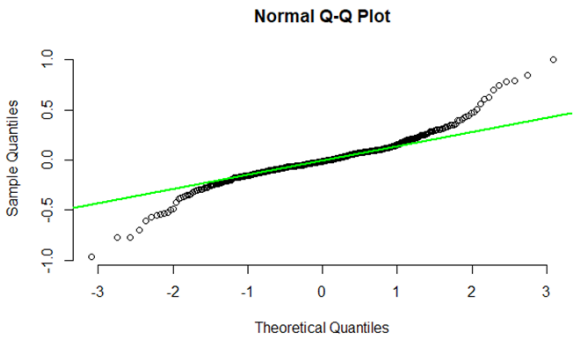


Figure 1.9: Q-Qplot of log-transformed model

Final Model

After appropriate model selection and log transformation, our final model of choice uses the logarithm of MEDV (Median Housing Prices) with covariates being LSTAT, RM, PTRATIO B, and CHAS. This model demonstrates a low degree of violation of the assumptions of linearity, normality, and homoscedasticity. It also shows a high adjusted R^2 of 0.73, which is an improvement over the previous model that did not use the log transformation. The relevant R output for the final model can be found in *Table 1.10*, and a more detailed discussion of the results is provided in the Discussion section.

	log(MEDV)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.17	2.83 – 3.51	**<0.001**
LSTAT	-0.03	-0.04 – -0.03	**<0.001**
RM	0.11	0.08 – 0.15	**<0.001**
PTRATIO	-0.04	-0.04 – -0.03	**<0.001**
B	0.00	0.00 – 0.00	**<0.001**
CHAS	0.13	0.06 – 0.20	**<0.001**
Observations	495		
R^2 / R^2 adjusted	0.738 / 0.735		

Table 1.10: Regression Result for the Final Model

Discussion and Conclusion

Our analysis investigated the relationship between several covariates and the median price of owner-occupied homes in Boston neighborhoods. After cleaning and preparing the dataset, our group employed exploratory data analysis, assessments of correlation, model selection techniques, and data transformation to identify significant variables and construct an effective model.

In our final model, we included five unique predictors along with a logarithmic transformation. Firstly, we have LSTAT (percentage of the population in each neighborhood considered lower income), which has a negative correlation with home prices. The coefficient for LSTAT is -0.0324, with a p-value of 0.000, indicating a strong and statistically significant relationship. This result is as expected, as lower-income areas are typically associated with more affordable housing. The number of rooms per household (RM) has a positive correlation with home prices, with a coefficient of 0.1125 and a p-value of 0.000, suggesting that more rooms are associated with higher housing costs, which intuitively makes sense. PTRATIO (the average number of students per teacher) has a negative coefficient of -0.0353, with a p-value of 0.000, indicating that as the number of students per teacher increases, housing costs decrease. This result is intuitive, as wealthier neighborhoods tend to have smaller student-to-teacher ratios. The variable B (proportion of Black residents) is also significant, with a coefficient of 0.0006 and a p-value of 0.000, reflecting that historical and socioeconomic factors, including race, could influence housing prices. Finally, CHAS (proximity to a river) has a strong positive correlation with housing prices, with a coefficient of 0.1313 and a p-value of 0.000, suggesting that natural features, such as proximity to rivers, could enhance the desirability and value of properties.

One main limitation of this analysis is the age and regional specificity of our dataset. Since the dataset is older (collected in 1970) and all the data is collected in Boston, this could limit the generalizability of our regression model when considering current housing markets or markets in other locations. Additionally, interactive models were not used since they would be hard to interpret in this context but by including them, our model may have improved in accuracy.

Finally, there are a few implications of our analysis. Understanding relationships between these covariates and housing prices provides insights for urban planners and policymakers. For example, enhancing school resources could uplift neighbourhood housing markets. Additionally, preserving and developing more natural features like the Charles River could also increase the housing desirability.

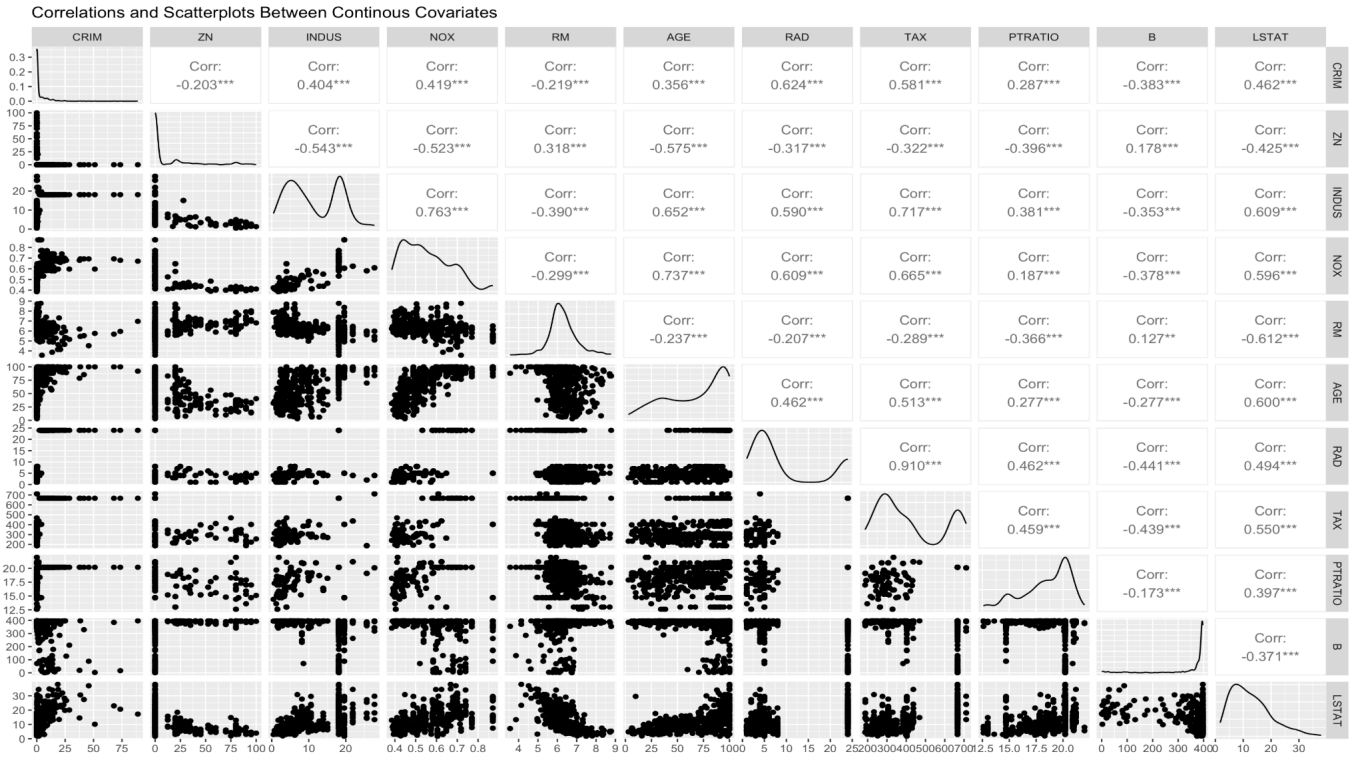
Overall, our analysis effectively demonstrates how the five given covariates can be used to create a regression model for Boston's housing price data. Future research could extend such findings by exploring additional variables that we did not consider or by using more recently collected data which could provide more context for how these relationships have changed over time.

Appendix

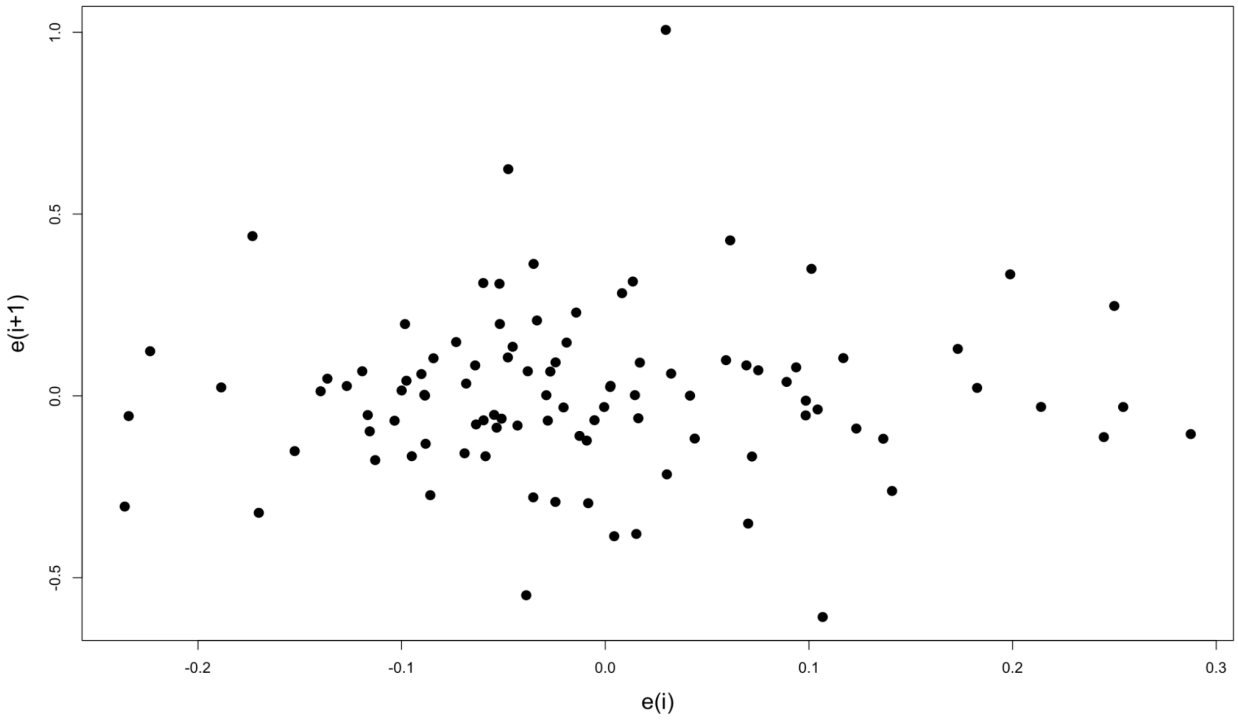
A. Summary statistics for all the variables

Variable	n	Mean	SD	Median	Min	Max
CRIM	495	3.69	8.68	0.28	0.01	88.98
ZN	495	11.45	23.55	0.00	0.00	100.00
INDUS	495	11.26	6.88	9.90	0.46	27.74
CHAS	495	0.07	0.26	0.00	0.00	1.00
NOX	495	0.56	0.12	0.54	0.38	0.87
RM	495	6.28	0.71	6.20	3.56	8.78
AGE	495	68.47	28.33	77.70	2.90	100.00
DIS	495	3.75	2.11	3.10	1.13	12.13
RAD	495	9.68	8.75	5.00	1.00	24.00
TAX	495	411.25	169.0	335.00	187.00	711.00
PTRATIO	495	18.50	2.16	19.10	12.60	22.00
B	495	355.8	92.13	391.27	0.32	396.90
LSTAT	495	12.67	7.11	11.38	1.73	37.97
MFRV	495	22.47	9.23	22.10	5.00	50.00

B. Pair-wise Correlation & Scatterplots for the full set of covariates



C. Serial residual plot between e_i and e_{i+1}



References

Ceccato, V., & Wilhelmsson, M. (2018). Does crime impact real estate prices? An assessment of accessibility and location. In G. J. N. Bruinsma & S. D. Johnson (Eds.), Oxford Handbook on Environmental Criminology (pp. 518-544). Oxford University Press.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic 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

Jones, C. A., Dunse, N. A., Li, E., & Liu, Y. (2023). Housing prices and the characteristics of nearby green space: Does landscape pattern index matter? *Land, 12*(2), 496. https://doi.org/10.3390/land12020496

StatLib (2016). boston [Data set]. StatLib. https://lib.stat.cmu.edu/datasets/boston