Introduction

Problem Statement — This project analyses the Boston Housing dataset to model and predict median residential property values (MEDV) from neighbourhood- and property-level features such as crime rate, proportion of residential land zoned for lots, average number of rooms per dwelling, the proportion of owner-occupied units built prior to 1940, distances to employment centres, property-tax rate, and air pollution. We build and compare predictive models to estimate house prices and identify the most influential variables.

Rationale — Housing prices are central to decisions made by homeowners, buyers, lenders, investors, and urban planners. Prices vary with location, neighbourhood characteristics, housing quality, and local amenities; understanding which features most strongly affect value helps stakeholders make informed pricing, investment, and policy decisions. Accurate predictive models also assist lenders with risk assessment and help policymakers target interventions to stabilise or improve local housing markets.

Application used — Jupyter Notebook (R), R Studio, VS Code, Google Colab and Github for remote repository.

Data Description

We have used the Boston Housing dataset originally compiled by Harrison and Rubinfeld (1978) and made widely available through the UCI Machine Learning Repository and various machine learning libraries. The dataset is accessible via R's `mlbench` module and also available on Kaggle and other data repositories. The data contains 506 observations (arranged as rows) representing different census tracts in the Boston area, and 14 variables (arranged as columns). The dataset includes 13 predictor variables (continuous and categorical) and 1 target variable (CMEDV) representing the median value of owner-occupied homes.

Variable	**Characteristic**	**Description**
CRIM	Continuous	Per capita crime rate by town
ZN	Continuous	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Continuous	Proportion of non-retail business acres per town
CHAS	Categorical - binary	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	Continuous	Nitric oxides concentration (parts per 10 million)
RM	Continuous	Average number of rooms per dwelling
AGE	Continuous	Proportion of owner-occupied units built prior to 1940
DIS	Continuous	Weighted distances to five Boston employment centres
RAD	Discrete	Index of accessibility to radial highways (1-24)
TAX	Continuous	Full-value property-tax rate per $10,000
PTRATIO	Continuous	Pupil-teacher ratio by town
B	Continuous	$1000(B_k - 0.63)^2$ where B_k is the proportion of Black residents by town
LSTAT	Continuous	Percentage of lower status of the population
CMEDV	Continuous	Median value of owner-occupied homes in $1000′s (target variable)

The dataset also includes fields such as town name, longitude, latitude, and census tract identifier. These fields are not used in the analysis as they do not contribute to the predictive modeling of house prices, but they are useful for geospatial analysis or mapping purposes.

Exploratory Data Analysis (EDA)

Summary Statistics — We begin with summary statistics for each variable, including minimum, first quartile, median, mean, third quartile, maximum, skewness, and kurtosis values. This provides an initial understanding of the data distribution, central tendency, potential outliers, and the shape of distributions.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Skewness	Kurtosis
cmedv	5.00	17.02	21.20	22.53	25.00	50.00	1.108	4.490
crim	0.00632	0.08205	0.25651	3.61352	3.67708	88.97620	5.208	39.753
zn	0.00	0.00	0.00	11.36	12.50	100.00	2.219	6.980
indus	0.46	5.19	9.69	11.14	18.10	27.74	0.294	1.767
nox	0.3850	0.4490	0.5380	0.5547	0.6240	0.8710	0.727	2.924
rm	3.561	5.886	6.208	6.285	6.623	8.780	0.402	4.861
age	2.90	45.02	77.50	68.57	94.08	100.00	−0.597	2.030
dis	1.130	2.100	3.207	3.795	5.188	12.127	1.009	3.471
rad	1.000	4.000	5.000	9.549	24.000	24.000	1.002	2.129
tax	187.0	279.0	330.0	408.2	666.0	711.0	0.668	1.857
ptratio	12.60	17.40	19.05	18.46	20.20	22.00	−0.800	2.706
b	0.32	375.38	391.44	356.67	396.23	396.90	−2.882	10.144
lstat	1.73	6.95	11.36	12.65	16.95	37.97	0.904	3.477

To better understand our dataset, we undertake univariate and bivariate analyses, visualising distributions and relationships between variables.

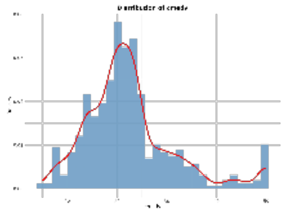


Figure 1: Distribution of CMEDV

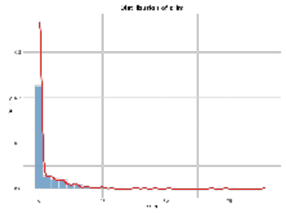


Figure 2: Distribution of CRIM

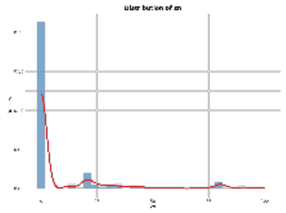


Figure 3: Distribution of ZN

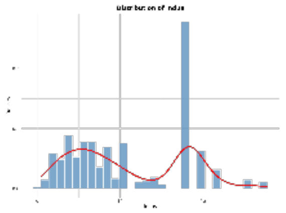


Figure 4: Distribution of INDUS

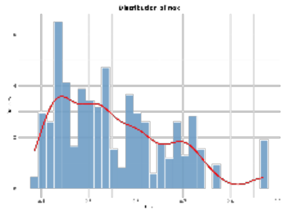


Figure 5: Distribution of NOX

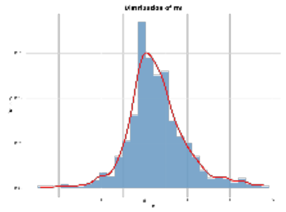


Figure 6: Distribution of RM

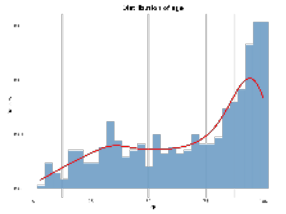


Figure 7: Distribution of AGE

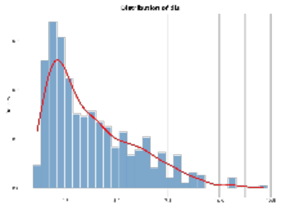
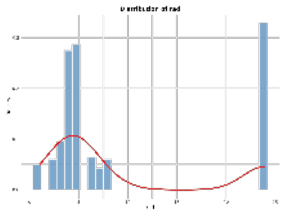
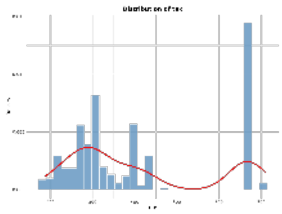
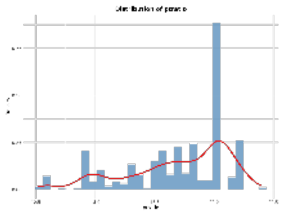
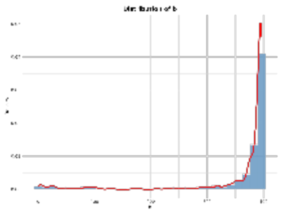
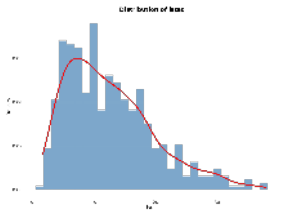
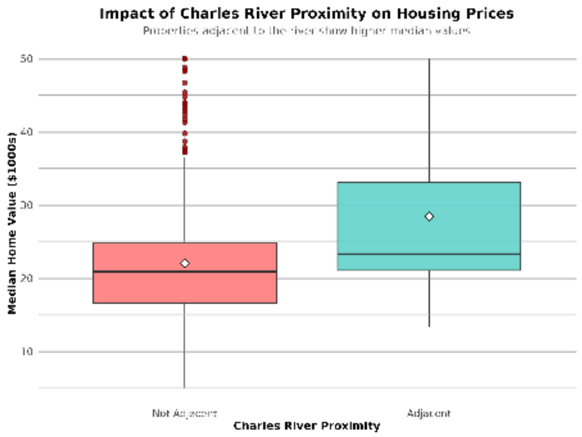
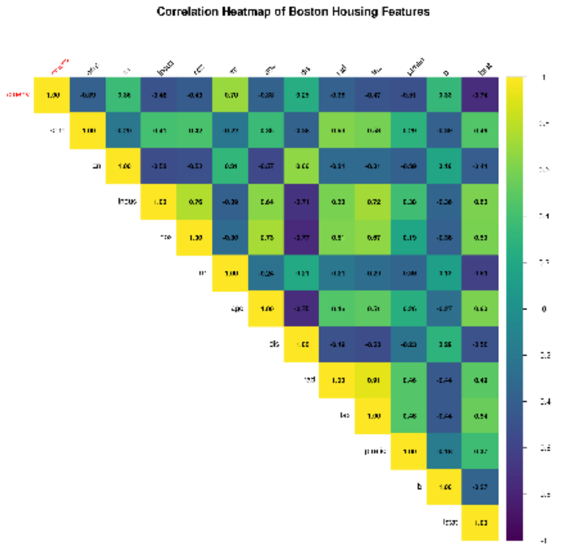
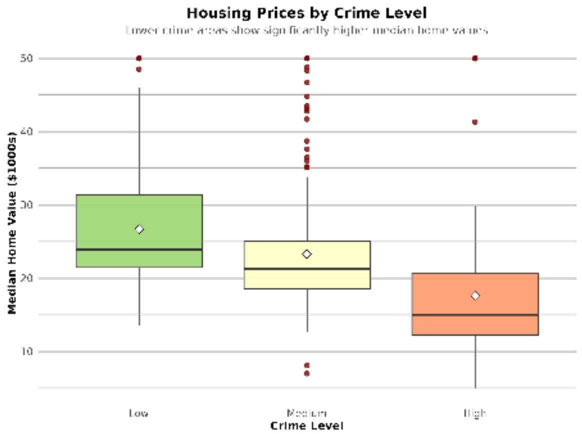


Figure 8: Distribution of DIS

 Figure 9: Distribution of RAD	 Figure 10: Distribution of TAX	 Figure 11: Distribution of PTRATIO	 Figure 12: Distribution of B
 Figure 13: Distribution of LSTAT			

Visualization	**Description**
 Figure 14: Properties by Charles River proximity	 Figure 15: Correlation heatmap
 Figure 16: CMEDV by crime level	

Geospatial Analysis

We visualize the geographic distribution of various features across the Boston area to understand spatial patterns and their relationship with housing prices.

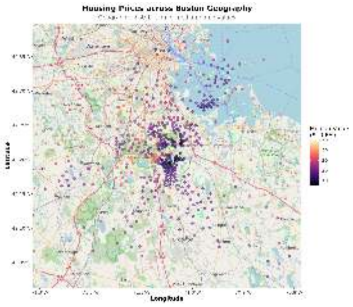


Figure 17: Geographic distribution of median home values

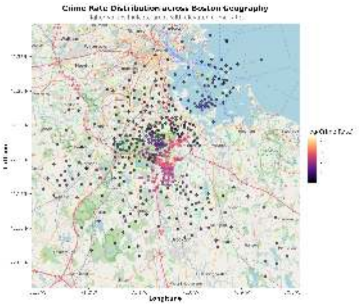


Figure 18: Geographic distribution of crime rate

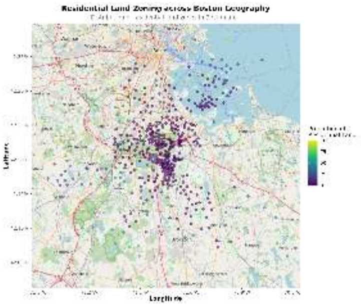


Figure 19: Geographic distribution of residential zoning

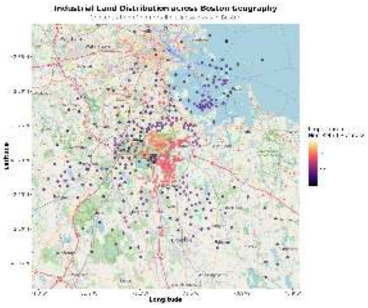


Figure 20: Geographic distribution of industrial areas

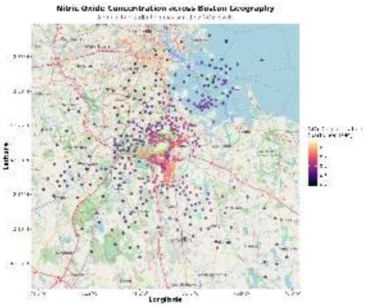


Figure 21: Geographic distribution of air pollution

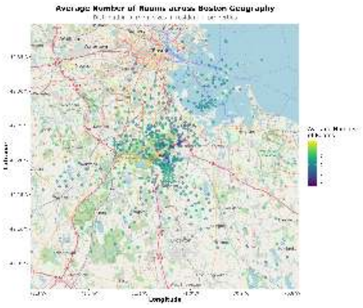


Figure 22: Geographic distribution of rooms per dwelling

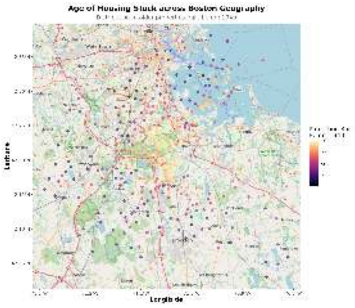


Figure 23: Geographic distribution of building age

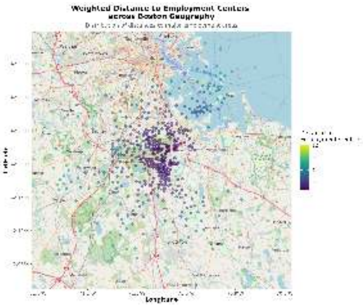

Weighted Distance to Employment Centers across Boston Geography
Latitude
Longitude

Figure 24: Geographic distribution of employment center distance

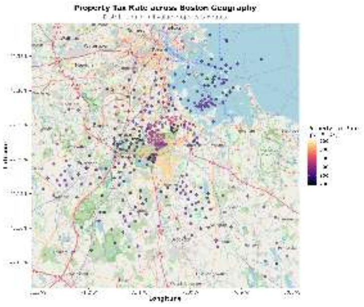


Figure 25: Geographic distribution of property tax rate

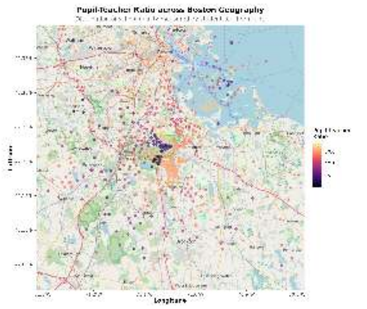


Figure 26: Geographic distribution of pupil-teacher ratio

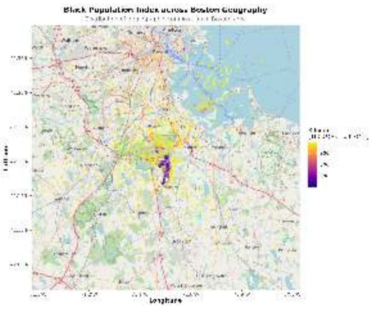


Figure 27: Geographic distribution of demographic variable

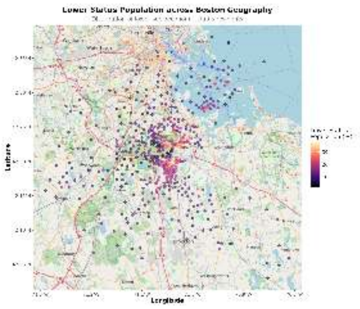


Figure 28: Geographic distribution of lower status population

Scatter Plot Analysis

We examine the relationship between each predictor variable and the target variable (CMEDV) through scatter plots.

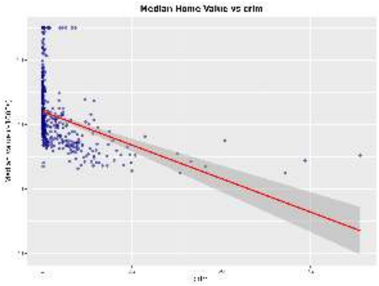


Figure 29: CMEDV vs CRIM

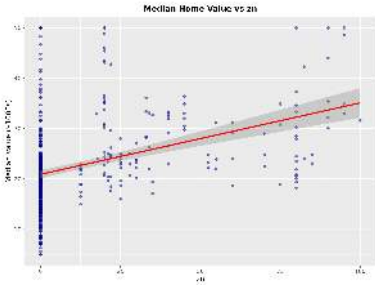


Figure 30: CMEDV vs ZN

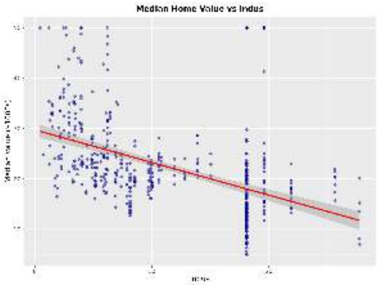


Figure 31: CMEDV vs INDUS

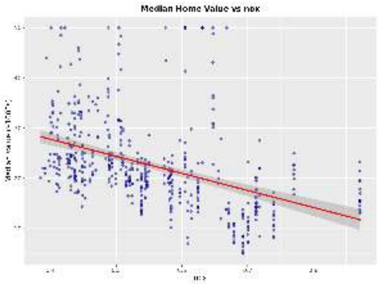


Figure 32: CMEDV vs NOX

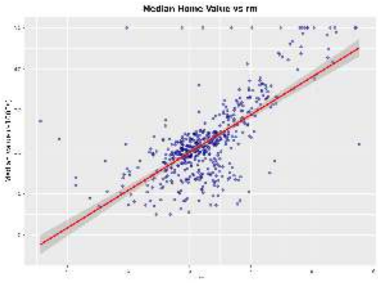


Figure 33: CMEDV vs RM

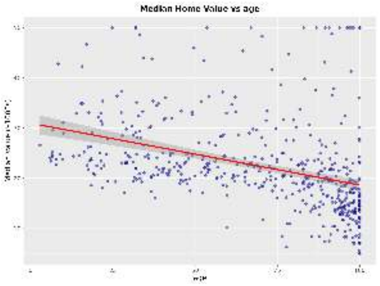


Figure 34: CMEDV vs AGE

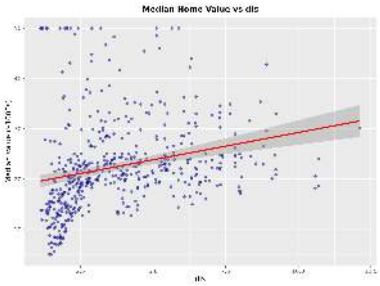


Figure 35: CMEDV vs DIS

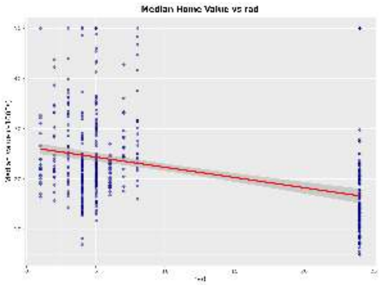


Figure 36: CMEDV vs RAD

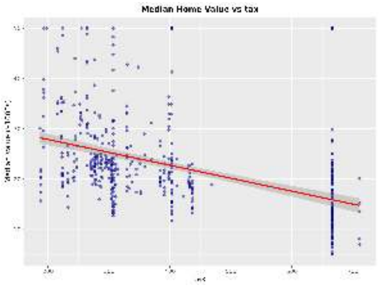


Figure 37: CMEDV vs TAX

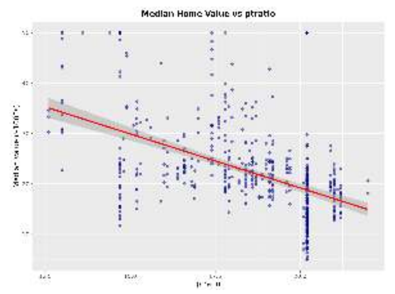


Figure 38: CMEDV vs PTRATIO

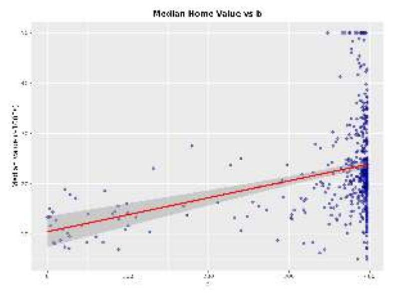


Figure 39: CMEDV vs B

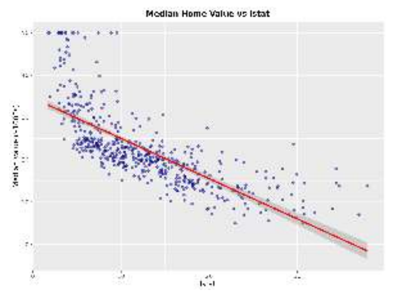


Figure 40: CMEDV vs LSTAT

Statistical Tests

Normality Testing — To assess whether our variables follow normal distributions, we apply the Shapiro-Wilk test to each continuous variable in the dataset. The Shapiro-Wilk test is a widely-used statistical test for normality, where the null hypothesis assumes the data is normally distributed. A p-value less than 0.05 indicates significant deviation from normality, leading us to reject the null hypothesis.

Variable	W Statistic	p-value	Normality
CMEDV	0.9169	4.63e-16	Not Normal
CRIM	0.4500	1.33e-36	Not Normal
ZN	0.5559	7.88e-34	Not Normal
INDUS	0.8998	1.06e-17	Not Normal
NOX	0.9356	5.78e-14	Not Normal
RM	0.9609	2.41e-10	Not Normal
AGE	0.8920	2.23e-18	Not Normal
DIS	0.9032	2.19e-17	Not Normal
RAD	0.6796	8.07e-30	Not Normal
TAX	0.8152	1.16e-23	Not Normal
PTRATIO	0.9036	2.36e-17	Not Normal
B	0.4768	6.06e-36	Not Normal
LSTAT	0.9369	8.29e-14	Not Normal

Interpretation — All variables in the Boston Housing dataset fail the Shapiro-Wilk normality test (p-value < 0.05), indicating significant departures from normal distribution. This finding has several implications for our analysis:

- **Variables with severe non-normality** (W < 0.7) include CRIM, ZN, RAD, and B, suggesting highly skewed distributions. These variables may require transformation (e.g., log or Box-Cox transformations) before applying parametric statistical methods.
- **Variables with moderate non-normality** (0.7 ≤ W < 0.9) include INDUS, AGE, DIS, TAX, and PTRATIO. While closer to normality, these still exhibit significant deviations and may benefit from transformation.
- **Variables closest to normality** (W ≥ 0.9) include NOX, RM, CMEDV, PTRATIO, DIS, and LSTAT. The target variable CMEDV (W = 0.9169) shows the least deviation among key variables, though it still significantly differs from a normal distribution.

- **Modeling implications** — The non-normality of predictor variables does not invalidate regression models (linear regression assumes normality of residuals, not predictors), but may affect interpretation and require robust statistical methods. Non-parametric approaches or transformations may improve model performance and meet distributional assumptions for certain statistical tests.
- **Outlier presence** — Low W statistics suggest the presence of outliers or heavy-tailed distributions, which aligns with our earlier observation of extreme values in variables like CRIM and B. These outliers warrant careful consideration during model development to avoid undue influence on predictions.

Correlation Significance Test — To determine which predictor variables have statistically meaningful relationships with median home value (CMEDV), we perform correlation significance tests. This analysis computes Pearson correlation coefficients and their associated p-values to identify which features exhibit reliable linear relationships with housing prices.

Variable	Correlation	P-value	Significance
LSTAT	−0.741	< 0.001	Significant
RM	0.696	< 0.001	Significant
PTRATIO	−0.506	< 0.001	Significant
INDUS	−0.485	< 0.001	Significant
TAX	−0.472	< 0.001	Significant
NOX	−0.429	< 0.001	Significant
CRIM	−0.390	< 0.001	Significant
RAD	−0.385	< 0.001	Significant
AGE	−0.378	< 0.001	Significant
ZN	0.360	< 0.001	Significant
B	0.335	< 0.001	Significant
DIS	0.249	< 0.001	Significant

Interpretation — All predictor variables exhibit statistically significant relationships with median home value ($p < 0.001$), indicating that each feature provides meaningful information for predicting housing prices. Key findings include:

- **Strongest negative predictor** — LSTAT (percentage of lower status population) shows the strongest negative correlation ($r = -0.741$), indicating that neighborhoods with higher proportions of lower-status residents have substantially lower median home values. This suggests socioeconomic factors are the most influential determinant of housing prices in the Boston area.
- **Strongest positive predictor** — RM (average number of rooms per dwelling) exhibits the strongest positive correlation ($r = 0.696$), demonstrating that larger homes with more rooms command significantly higher prices. This reflects the premium placed on living space and home size in property valuations.
- **Moderate negative predictors** — PTRATIO ($r = -0.506$), INDUS ($r = -0.485$), and TAX ($r = -0.472$) show moderate negative correlations, suggesting that neighborhoods with higher pupil-teacher ratios, more industrial development, and higher property taxes tend to have lower home values. These factors likely reflect less desirable residential environments or higher cost burdens on homeowners.

- **Environmental and accessibility factors** — NOX (air pollution, r = −0.429), RAD (highway accessibility, r = −0.385), and DIS (distance to employment centers, r = 0.249) demonstrate that environmental quality and geographic location play significant roles in housing values. Surprisingly, greater highway accessibility correlates negatively with prices, possibly due to noise and congestion effects.
- **Neighborhood characteristics** — CRIM (crime rate, r = −0.390) and AGE (proportion of older homes, r = −0.378) both negatively affect home values, confirming that safety concerns and building age reduce property desirability. Meanwhile, ZN (residential zoning, r = 0.360) and B (demographic index, r = 0.335) show positive but more modest associations with prices.
- **Modeling implications** — The strong correlations of LSTAT and RM suggest these variables should be prioritized in predictive models. The presence of multiple significant but moderately correlated predictors indicates that multivariate regression models incorporating several features will likely outperform simple univariate models. However, the correlation structure also raises concerns about potential multicollinearity between predictors (e.g., TAX and RAD, INDUS and NOX), which may need to be addressed during model development through variable selection or regularization techniques.

ANOVA: Charles River Impact Analysis — To test whether proximity to the Charles River has a statistically significant effect on housing prices, we perform a one-way Analysis of Variance (ANOVA). This test compares mean home values between properties adjacent to the river (CHAS = 1) and those not adjacent (CHAS = 0).

Source	**Df**	**Sum Sq**	**Mean Sq**	**F-value**	**p-value**
Charles River (CHAS)	1	1314	1313.8	16.05	7.11e-05 ***
Residuals	504	41264	81.9		

Interpretation — The ANOVA results provide strong statistical evidence that proximity to the Charles River significantly affects median home values:

- **Statistical significance** — The F-statistic of 16.05 with a p-value of 0.00007 ($p < 0.001$) indicates we can reject the null hypothesis that river proximity has no effect on housing prices. This result is highly statistically significant, marked with three asterisks (***), meaning there is less than 0.01% probability that the observed difference occurred by chance.
- **Effect magnitude** — Properties adjacent to the Charles River have a mean value of $28,440, compared to $22,090 for properties not adjacent to the river. This represents a difference of approximately $6,350, or a 29% premium for river-adjacent properties. This substantial price differential reflects the desirability of waterfront locations and scenic views.
- **Variance explained** — The Charles River variable accounts for 1314 units of sum of squares out of the total variance, representing approximately 3.1% of the total variation in housing prices. While statistically significant, this suggests that river proximity is one of many factors influencing home values, consistent with our earlier correlation analysis showing multiple significant predictors.
- **Practical implications** — For homebuyers, real estate investors, and urban planners, this finding confirms that waterfront location commands a measurable premium in the Boston housing market. The consistent and significant effect suggests this premium is not merely a result of confounding factors, but represents genuine value placed on river proximity by the housing market.

- **Model considerations** — The CHAS variable should be retained as a predictor in our regression models given its significant relationship with housing prices. As a binary categorical variable, it captures a distinct geographic feature that cannot be adequately represented by other continuous predictors in the dataset.

Key Insights from Exploratory Analysis

Through our comprehensive exploratory data analysis, we have identified several critical patterns and relationships that provide valuable insights into the factors driving housing prices in the Boston area:

Target Variable Distribution — The median home values (CMEDV) range from $5,000 to $50,000, with a mean of $22,530 and a median of $21,200. The distribution shows positive skewness (1.108), indicating a concentration of lower-priced homes with a tail of higher-valued properties. The maximum value of $50,000 likely represents censored data, as several observations cluster at this ceiling, suggesting some properties may have been worth more but were capped at this reporting threshold.

Primary Value Driver: Living Space — The average number of rooms per dwelling (RM) emerges as the strongest positive predictor of home value (r = 0.696). This finding underscores that property size and living space are paramount considerations for homebuyers in the Boston market. Each additional room corresponds to a substantial increase in median home value, reflecting the premium buyers place on larger, more spacious residences. This relationship holds consistently across the dataset and should be a central feature in any predictive model.

Primary Value Detractor: Socioeconomic Status — The percentage of lower-status population (LSTAT) exhibits the strongest negative correlation with home values (r = −0.741), making it the single most influential predictor variable. This powerful inverse relationship indicates that neighborhoods with higher concentrations of lower-income residents experience substantially depressed property values. This finding highlights how socioeconomic stratification directly translates to housing market outcomes, with implications for housing affordability, neighborhood development, and urban inequality.

Data Quality Challenge: Crime Rate Skewness — The per capita crime rate (CRIM) shows extreme positive skewness (5.208) with a maximum value of 88.98 compared to a mean of 3.61. This severe right-skewed distribution indicates that while most neighborhoods experience relatively low crime rates, a small number of census tracts suffer from exceptionally high crime levels. This skewness presents modeling challenges and suggests that log transformation or other normalization techniques may be necessary to prevent these extreme values from disproportionately influencing regression models.

Anomalous Data Pattern: Demographic Variable — The B variable (demographic index related to the proportion of Black residents) contains the most outliers in the dataset, with an extremely high kurtosis value (10.144) and severe negative skewness (-2.882). The formula $1000(B_k - 0.63)^2$ produces values ranging from 0.32 to 396.90, with most observations clustering near the maximum. This unusual distribution and the variable's historical context warrant careful consideration during modeling, as the extreme concentration of values may limit its predictive utility and raises questions about measurement methodology and data quality in this particular feature.

Synthesis — These five key insights collectively paint a picture of Boston's housing market as one driven primarily by property characteristics (room count) and neighborhood socioeconomic composition (lower-status population percentage), while also revealing data quality issues (crime skewness and demographic variable outliers) that must be addressed in subsequent modeling phases.

The interplay between these factors—particularly the dominant roles of RM and LSTAT—will be central to developing accurate and interpretable predictive models for median home values.

Regression Modeling

We build a Multiple Linear Regression (MLR) model using all 13 predictor variables to predict median home values (CMEDV). The model is trained on 407 observations from the training dataset.

Model Equation — The fitted regression plane is expressed as:

$$003 \times \text{INDUS} + 2.378 \times \text{CHAS} - 16.773 \times \text{NOX} + 3.636 \times \text{RM} - 0.005 \times \text{AGE} - 1.505 \times \text{DIS} + 0.317 \times \text{RAD} - 0.013 \times \text{TA}$$

Model Performance — The multiple linear regression model achieves strong predictive performance:

- **R-squared** — The model explains 73.81% of the variance in median home values, indicating that nearly three-quarters of the variation in housing prices can be accounted for by the 13 predictor variables. The adjusted R-squared of 72.94% confirms this result after penalizing for the number of predictors, suggesting minimal overfitting.
- **Overall significance** — The F-statistic of 85.19 with p-value < 2.2e-16 demonstrates that the model as a whole is highly statistically significant, meaning at least one predictor has a non-zero effect on home values.
- **Residual standard error** — The model's prediction errors have a standard deviation of $4,834, representing approximately 21% of the mean home value ($22,530), which indicates reasonable but not perfect prediction accuracy.

Coefficient Interpretation and Significance —

- **Highly significant predictors (p < 0.001)** — LSTAT (t = −9.411), RM (t = 7.943), NOX (t = −3.907), DIS (t = −6.560), PTRATIO (t = −6.443), and the intercept all show extremely strong statistical significance. These variables are the most reliable predictors in the model.
- **Moderately significant predictors (0.001 ≤ p < 0.05)** — CRIM (p = 0.022), ZN (p = 0.016), CHAS (p = 0.015), RAD (p < 0.001), TAX (p = 0.003), and B (p = 0.001) contribute meaningfully to the model, though with smaller effect sizes or higher variability.
- **Non-significant predictors** — INDUS (p = 0.961) and AGE (p = 0.741) show no statistically significant relationship with home values when controlling for other variables. These variables add little predictive value and could potentially be removed to simplify the model.

Key Predictor Effects —

- **Largest positive effect** — Each additional room (RM) increases median home value by $3,636, the largest positive coefficient in the model. This confirms our earlier finding that living space is the primary driver of higher housing prices.
- **Largest negative effect** — Each 1% increase in lower-status population (LSTAT) decreases median home value by $543. Combined with its high statistical significance (t = −9.411), LSTAT remains the strongest predictor of lower home values.
- **Environmental quality** — NOX (nitric oxide concentration) has the second-largest coefficient magnitude at -$16,773 per unit increase, indicating that air pollution severely depresses home values. This substantial effect underscores the importance of environmental quality in housing markets.

- **School quality** — Each unit increase in pupil-teacher ratio (PTRATIO) reduces home values by $963, reflecting the premium buyers place on better-funded schools with smaller class sizes.
- **Waterfront premium** — Properties adjacent to the Charles River (CHAS = 1) command a $2,378 premium, consistent with our ANOVA findings.

Residual Analysis — The residuals range from -$15,256 to $25,659, with the distribution showing:

- The median residual (-$504) is close to zero, suggesting generally unbiased predictions.
- The interquartile range spans from -$2,808 to $1,597, indicating most predictions are within approximately $2,200 of actual values.
- The maximum positive residual of $25,659 suggests the model significantly underestimates some high-value properties, possibly those at or near the $50,000 ceiling.
- Asymmetry in the residual range (larger positive outliers) indicates potential heteroscedasticity or non-normality in the error distribution.

Model Limitations and Considerations —

- The non-significance of INDUS and AGE suggests potential multicollinearity, where these variables' effects may be captured by other correlated predictors. Variance inflation factor (VIF) analysis would help identify problematic collinearity.
- The presence of large residuals, particularly the maximum of $25,659, indicates the linear model may not adequately capture non-linear relationships or interaction effects between predictors.
- Variables requiring potential transformation (based on earlier normality tests) were used in their original scale, which may limit model performance. Log or polynomial transformations of skewed variables like CRIM could improve fit.

Multicollinearity Assessment Using Variance Inflation Factor (VIF) — To investigate potential multicollinearity among predictor variables, we calculate Variance Inflation Factor (VIF) values for each variable in the multiple linear regression model. VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. A VIF value of 1 indicates no correlation with other predictors, while values exceeding 5 or 10 suggest moderate or severe multicollinearity, respectively.

Variable	VIF	Status
TAX	9.132	MODERATE
RAD	7.668	MODERATE
NOX	4.370	LOW
INDUS	4.024	LOW
DIS	4.002	LOW
AGE	3.165	LOW
LSTAT	3.004	LOW
ZN	2.385	LOW
RM	1.916	LOW
PTRATIO	1.848	LOW
CRIM	1.847	LOW
B	1.387	LOW
CHAS	1.064	LOW

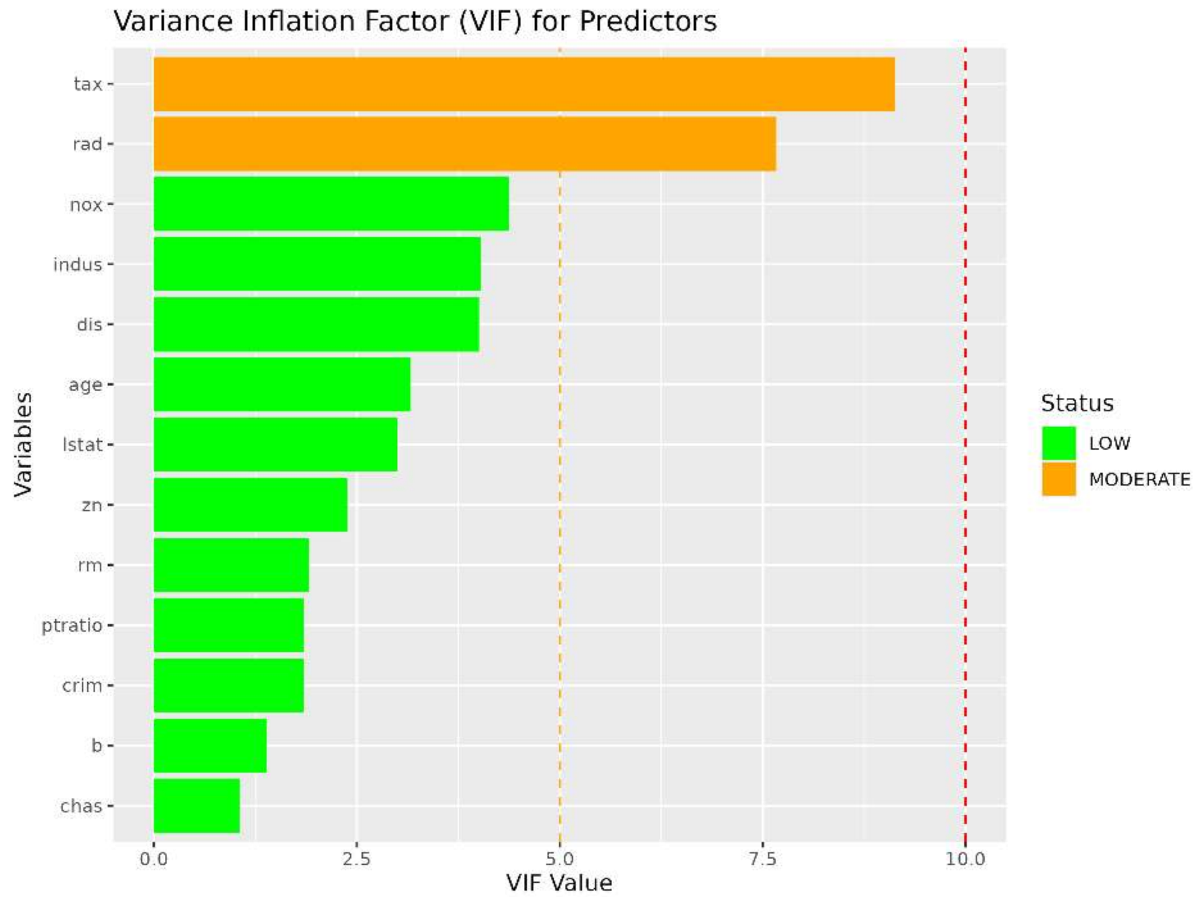


Figure 41: Variance Inflation Factor (VIF) for predictor variables

Interpretation — The VIF analysis reveals moderate multicollinearity concerns for two variables while confirming that most predictors maintain acceptable levels of independence:

- **Moderate multicollinearity** — TAX (property tax rate, VIF = 9.132) and RAD (highway accessibility index, VIF = 7.668) exhibit elevated VIF values approaching the threshold of concern. These two variables likely share substantial variance, which is logical given that property tax rates often correlate with infrastructure development and highway access. While not severe enough to invalidate the model, this moderate collinearity may inflate the standard errors of their coefficient estimates and reduce their individual statistical power.
- **Acceptable collinearity levels** — The remaining 11 predictor variables all show VIF values below 5, indicating low multicollinearity. Notable findings include: NOX (VIF = 4.370), INDUS (VIF = 4.024), and DIS (VIF = 4.002) approach but remain below the moderate threshold, while strongly predictive variables like LSTAT (VIF = 3.004) and RM (VIF = 1.916) demonstrate excellent independence from other predictors.
- **Explaining non-significant coefficients** — The earlier finding that INDUS and AGE were not statistically significant in the regression model is partially explained by their moderate VIF values (4.024 and 3.165, respectively). While not severely collinear, these variables share enough variance with other predictors that their unique contributions to the model become statistically indistinguishable from zero. This suggests their effects may be captured by correlated variables like NOX, DIS, or TAX.
- **TAX and RAD relationship** — The moderate collinearity between TAX and RAD makes conceptual sense: census tracts with better highway access (higher RAD) often have higher

property values and consequently higher tax assessments, or alternatively, areas with extensive infrastructure development face higher tax burdens to fund maintenance. This shared variance complicates interpretation of their individual effects—the negative coefficients for both variables in the regression may partially reflect their joint influence rather than independent effects.

- **Model stability implications** — Despite the moderate collinearity in TAX and RAD, the overall VIF profile is acceptable for the multiple linear regression model. No variables exceed the critical threshold of VIF = 10, meaning the model's coefficient estimates remain reasonably stable and interpretable. However, if we were to pursue variable selection or model simplification, removing one variable from the TAX-RAD pair could improve model parsimony without substantial loss of predictive power.
- **Recommendation for model refinement** — Consider creating a simplified model that either (1) removes one of the correlated pair (TAX or RAD) to reduce collinearity, or (2) combines them into a composite "neighborhood infrastructure" index. Additionally, given their non-significance and moderate VIF values, INDUS and AGE are candidates for removal in a parsimonious model, which could improve interpretability without sacrificing predictive accuracy.

Refined Model: Addressing Multicollinearity — Based on the VIF analysis revealing moderate multicollinearity in TAX (VIF = 9.132) and RAD (VIF = 7.668), we develop a refined model by removing these correlated variables. The simplified model retains 11 predictors, eliminating the two variables with elevated VIF values while preserving all other features.

Refined Model Equation — The updated regression model is:

$$0.034 \times \text{ZN} - 0.057 \times \text{INDUS} + 2.783 \times \text{CHAS} - 13.615 \times \text{NOX} + 3.904 \times \text{RM} - 0.009 \times \text{AGE} - 1.519 \times \text{DIS} - 0.817 \times \text{PT}$$

Multicollinearity Improvement — The refined model successfully reduces multicollinearity across all predictors:

- All VIF values now fall below 5, with the highest being DIS (VIF = 4.001), NOX (VIF = 3.904), and INDUS (VIF = 3.322). This represents a substantial improvement over the original model where TAX and RAD exhibited VIF values exceeding 7.
- Most predictors show VIF values between 1 and 3.5, indicating minimal correlation with other variables and ensuring stable, interpretable coefficient estimates.
- The removal of TAX and RAD has not introduced new collinearity issues, confirming these variables were the primary source of multicollinearity concerns.

Model Performance Comparison —

Metric	Original Model	Refined Model	Change
R-squared	0.7381	0.7261	−0.012
Adjusted R-squared	0.7294	0.7185	−0.0109
Residual Std. Error	$4,834	$4,930	+$96
F-statistic	85.19	95.19	+10.0
Degrees of Freedom (predictors)	13	11	−2

The refined model shows minimal degradation in explanatory power despite removing two variables:

- **Modest R-squared reduction** — The R-squared decreased from 73.81% to 72.61%, a loss of only 1.2 percentage points. This indicates that TAX and RAD contributed relatively little unique predictive information beyond what other variables already captured.

- **Adjusted R-squared nearly unchanged** — The adjusted R-squared (which penalizes model complexity) decreased from 72.94% to 71.85%, a difference of just 1.09 percentage points. This minimal change suggests the simpler model achieves nearly equivalent performance with better parsimony.
- **Slight increase in prediction error** — The residual standard error increased marginally from $4,834 to $4,930, representing approximately $96 in additional average prediction error. This small increase is acceptable given the substantial reduction in multicollinearity.
- **Improved F-statistic** — The F-statistic increased from 85.19 to 95.19, indicating stronger overall model significance. This counterintuitive improvement occurs because the F-statistic accounts for model complexity, and the refined model achieves similar explanatory power with fewer predictors.

Changes in Coefficient Significance —

- **Newly significant variables** — ZN (p = 0.033) and B (p = 0.003) became statistically significant in the refined model, whereas they showed borderline significance in the original model. Removing the correlated TAX and RAD variables allowed these predictors' unique contributions to emerge more clearly.
- **Maintained significance** — CHAS, NOX, RM, DIS, PTRATIO, and LSTAT remain highly significant (p < 0.01), confirming their robust predictive value regardless of model specification.
- **Persistent non-significance** — CRIM (p = 0.285), INDUS (p = 0.377), and AGE (p = 0.557) remain statistically non-significant. These variables appear to contribute little unique information when controlling for other predictors and could be candidates for further model simplification.

Coefficient Magnitude Changes —

- **RM effect increase** — The coefficient for RM increased from $3,636 to $3,904, suggesting that removing TAX and RAD (which may have partially captured neighborhood quality) allowed the rooms variable to express a slightly stronger relationship with home values.
- **NOX effect reduction** — The NOX coefficient changed from -$16,773 to -$13,615, a decrease in magnitude. This suggests that some of the air pollution effect in the original model was confounded with the removed variables, particularly RAD (highway access) which often correlates with pollution levels.
- **LSTAT coefficient unchanged** — The LSTAT coefficient remained virtually identical (-$543 vs. -$541), confirming its independent and robust relationship with housing prices.

Model Selection Justification — The refined model represents a superior balance between predictive accuracy and model interpretability:

- By eliminating TAX and RAD, we have addressed the primary multicollinearity concerns without sacrificing meaningful explanatory power (less than 2% reduction in adjusted R-squared).
- The improved VIF profile ensures coefficient estimates are more stable and reliable, reducing the risk of coefficient sign reversals or magnitude distortions due to collinearity.
- The simpler model with 11 predictors is easier to interpret and implement, while the increased F-statistic confirms it achieves strong overall significance.
- The persistence of non-significant variables (CRIM, INDUS, AGE) suggests potential for further refinement, though their removal would require careful consideration of theoretical importance versus statistical parsimony.

Coefficient Summary and Statistical Inference — The refined model's coefficient estimates, standard errors, t-values, and confidence intervals provide detailed insights into the magnitude, precision, and reliability of each predictor's effect on median home values:

Variable	Estimate	Std. Error	t-value	p-value	Significance
(Intercept)	31.0382	5.5220	5.621	3.60e-08	***
CRIM	−0.0392	0.0366	−1.071	2.85e-01	
ZN	0.0335	0.0157	2.143	3.27e-02	*
INDUS	−0.0575	0.0649	−0.885	3.77e-01	
CHAS	2.7831	0.9872	2.819	5.06e-03	**
NOX	−13.6147	4.1392	−3.289	1.09e-03	**
RM	3.9045	0.4618	8.455	5.44e-16	***
AGE	−0.0093	0.0158	−0.589	5.56e-01	
DIS	−1.5185	0.2340	−6.489	2.59e-10	***
PTRATIO	−0.8172	0.1395	−5.860	9.77e-09	***
B	0.0088	0.0029	2.995	2.92e-03	**
LSTAT	−0.5410	0.0588	−9.194	2.20e-18	***

95% Confidence Intervals for Coefficients — The confidence intervals provide ranges within which we can be 95% confident the true population coefficients lie:

Variable	Lower Bound (2.5%)	Upper Bound (97.5%)
(Intercept)	20.1819	41.8944
CRIM	−0.1110	0.0327
ZN	0.0028	0.0643
INDUS	−0.1851	0.0702
CHAS	0.8422	4.7240
NOX	−21.7524	−5.4770
RM	2.9966	4.8123
AGE	−0.0404	0.0218
DIS	−1.9786	−1.0585
PTRATIO	−1.0914	−0.5430
B	0.0030	0.0146
LSTAT	−0.6567	−0.4253

Inference and Interpretation —

- **Model intercept** — The baseline median home value (when all predictors equal zero) is estimated at $31,038 with extremely high statistical significance (t = 5.621, p < 0.001). The 95% confidence interval [$20,182, $41,894] indicates we can be highly confident this baseline value falls within this range, though this interpretation has limited practical meaning since many predictor values cannot realistically equal zero.
- **Most influential predictor: LSTAT** — Each 1% increase in lower-status population decreases median home value by $541 (95% CI: [-$657, -$425]). With the largest absolute t-value (-9.194) and smallest p-value (2.20e-18), LSTAT demonstrates the strongest and most reliable relationship with housing prices. The narrow confidence interval confirms high precision in this estimate, making it the most stable predictor in the model.

- **Second most influential predictor: RM** — Each additional room increases median home value by $3,905 (95% CI: [$2,997, $4,812]). The exceptionally high t-value (8.455) and tiny p-value (5.44e-16) establish RM as the second-strongest predictor. The confidence interval, while wider than LSTAT's due to the larger coefficient magnitude, remains entirely positive and precise, confirming the robust positive effect of living space on property values.
- **Distance to employment centers: DIS** — Each unit increase in weighted distance to employment centers decreases median home value by $1,519 (95% CI: [-$1,979, -$1,059]). The large negative t-value (-6.489) and highly significant p-value (2.59e-10) indicate that proximity to job centers is a critical determinant of housing prices. The entirely negative confidence interval confirms this effect is reliably detrimental to home values.
- **School quality: PTRATIO** — Each unit increase in pupil-teacher ratio reduces median home value by $817 (95% CI: [-$1,091, -$543]). With a t-value of −5.860 and p-value of 9.77e-09, this demonstrates that neighborhoods with better-funded schools (lower pupil-teacher ratios) command significantly higher property values. The confidence interval confirms this negative effect with high certainty.
- **Air pollution: NOX** — Each unit increase in nitric oxide concentration decreases median home value by $13,615 (95% CI: [-$21,752, -$5,477]). While highly significant (t = −3.289, p = 0.001), the wide confidence interval reflects greater uncertainty in this estimate, likely due to the variable's complex relationship with other urban characteristics. Nonetheless, the interval remains entirely negative, confirming environmental quality's importance.
- **Charles River proximity: CHAS** — Properties adjacent to the Charles River command a premium of $2,783 (95% CI: [$842, $4,724]). This effect is statistically significant (t = 2.819, p = 0.005), though the relatively wide confidence interval indicates moderate uncertainty in the precise premium magnitude. The entirely positive interval confirms that waterfront location consistently adds value.
- **Demographic variable: B** — Each unit increase in the B index raises median home value by $8.80 (95% CI: [$3.00, $14.60]). While statistically significant (t = 2.995, p = 0.003), the small coefficient magnitude means this variable has minimal practical impact on housing prices. The narrow but entirely positive confidence interval suggests a consistent but modest effect.
- **Residential zoning: ZN** — Each percentage point increase in residential land zoned for large lots increases median home value by $33.50 (95% CI: [$2.80, $64.30]). This effect is marginally significant (t = 2.143, p = 0.033), and the confidence interval—while entirely positive—is relatively wide and includes values near zero, indicating lower precision and suggesting this effect may be weaker or less consistent across the dataset.
- **Non-significant predictors** — CRIM (95% CI: [-$111, $33]), INDUS (95% CI: [-$185, $70]), and AGE (95% CI: [-$40, $22]) all have confidence intervals that include zero, confirming their lack of statistical significance. These variables contribute little unique predictive information when controlling for other features, as their effects cannot be reliably distinguished from zero. The confidence intervals spanning both positive and negative values indicate substantial uncertainty about even the direction of their relationships with housing prices.
- **Overall model significance** — The F-statistic of 95.19 with p-value of 8.72e-104 provides overwhelming evidence that the refined model as a whole is highly statistically significant. This extremely small p-value (essentially zero) indicates that the probability of observing such strong relationships by chance alone is infinitesimally small, confirming that the predictor variables collectively provide substantial explanatory power for median home values.

Key Takeaways — The coefficient analysis reveals that six variables (LSTAT, RM, DIS, PTRATIO, NOX, and CHAS) exhibit strong, reliable, and precisely estimated effects on housing prices, while three variables (CRIM, INDUS, and AGE) contribute negligibly and could be removed to further simplify the model. The magnitude of coefficient estimates, combined with their confidence intervals, provides quantitative guidance for understanding housing market dynamics: living space and socioeconomic composition dominate price determination, while employment accessibility, school quality, environmental conditions, and waterfront location also play significant roles.

Regression Diagnostics — To assess the validity of our multiple linear regression model assumptions, we examine four key diagnostic plots that evaluate linearity, homoscedasticity, normality of residuals, and influence of individual observations.

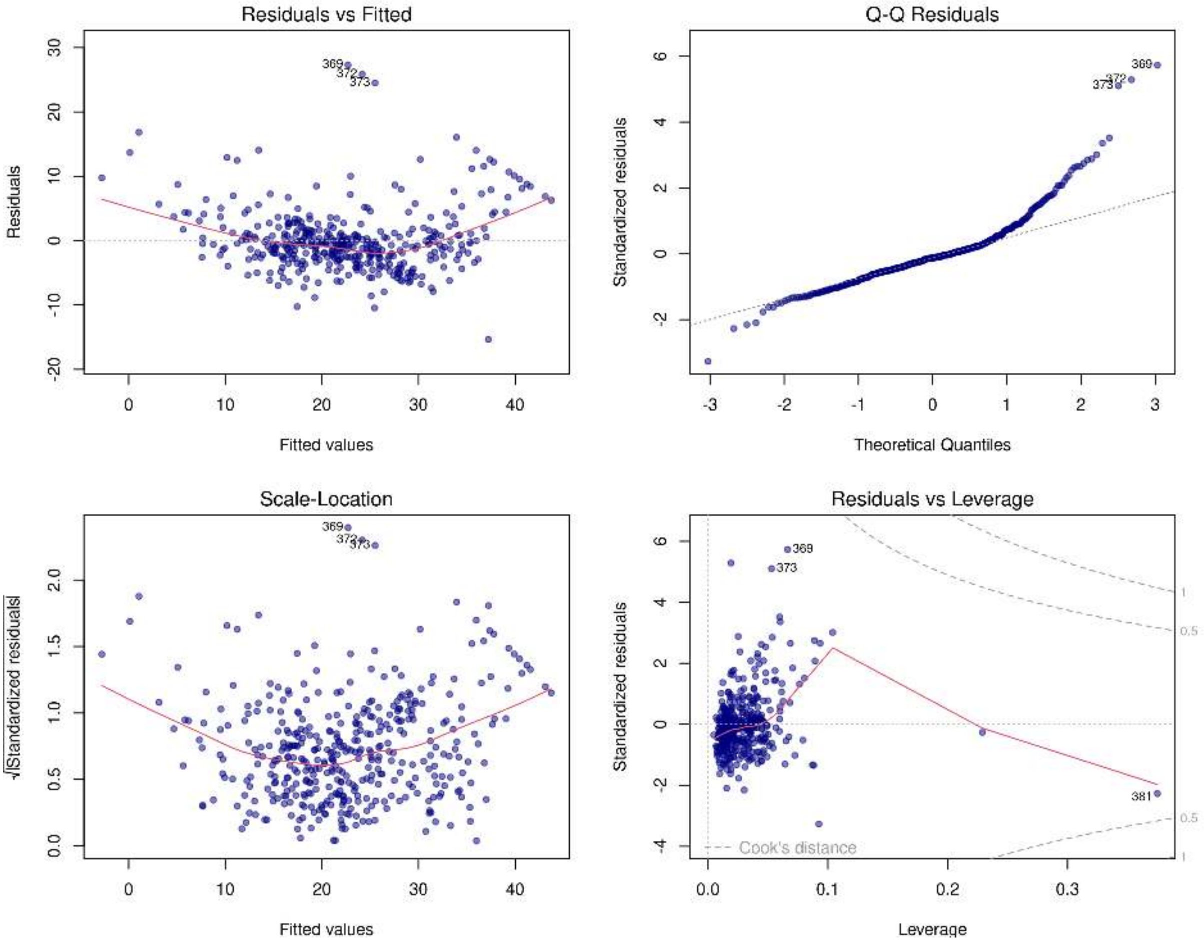


Figure 42: Regression diagnostic plots for model validation

The diagnostic plots reveal that while the model performs reasonably well, there are some deviations from ideal linear regression assumptions. The Residuals vs Fitted plot shows a relatively random scatter around the zero line, suggesting adequate linearity, though some heteroscedasticity may be present with slightly increasing variance at higher fitted values. The Q-Q plot indicates that residuals follow an approximately normal distribution in the central range, but with notable deviations in the tails, particularly for extreme observations. The Scale-Location plot confirms mild heteroscedasticity, as the spread of standardized residuals increases somewhat with fitted values. The Residuals vs Leverage plot identifies several high-leverage observations (notably observations 369, 381, and 373), though none appear to be overly influential outliers with high Cook's distance, suggesting the model's coefficient estimates remain stable despite these data points.

Residual Assumption Testing — Beyond visual diagnostics, we perform formal statistical tests to rigorously assess whether our regression model satisfies the critical assumptions of homoscedasticity (constant variance) and independence of residuals.

Breusch-Pagan Test for Heteroscedasticity — The Breusch-Pagan test evaluates whether the variance of residuals remains constant across all levels of the predictor variables (homoscedasticity) or varies systematically (heteroscedasticity). The test statistic follows a chi-squared distribution, and the null hypothesis assumes constant variance.

Test	BP Statistic	p-value	Result
Breusch-Pagan	43.535 (df = 11)	8.763e-06	Heteroscedasticity Detected

Interpretation — The Breusch-Pagan test yields a test statistic of BP = 43.535 with 11 degrees of freedom and an extremely small p-value of 8.763e-06 (p < 0.001). This result provides strong statistical evidence to reject the null hypothesis of homoscedasticity, confirming that the variance of residuals is not constant across the range of fitted values. This finding aligns with our earlier visual observation in the Scale-Location diagnostic plot, where we noted increasing variance at higher predicted home values.

The presence of heteroscedasticity has several important implications: First, while coefficient estimates remain unbiased, their standard errors become unreliable, potentially leading to incorrect inference about statistical significance. Second, confidence intervals and hypothesis tests may be invalid, as they assume constant variance. Third, the model's efficiency is reduced—predictions for certain ranges of home values will be more uncertain than others.

To address this violation, we could consider: (1) applying variance-stabilizing transformations such as log transformation to the dependent variable (CMEDV), (2) using weighted least squares regression to give less weight to observations with higher variance, or (3) computing heteroscedasticity-robust standard errors (White's correction) to obtain valid inference despite non-constant variance.

Durbin-Watson Test for Autocorrelation — The Durbin-Watson test examines whether residuals are independent or exhibit serial correlation (autocorrelation). The test statistic ranges from 0 to 4, where values near 2 indicate no autocorrelation, values below 1.5 suggest positive autocorrelation, and values above 2.5 indicate negative autocorrelation.

Test	Lag	Autocorrelation	D-W Statistic	Result
Durbin-Watson	1	0.4132	1.166	Positive Autocorrelation

Interpretation — The Durbin-Watson test produces a statistic of DW = 1.166 with an autocorrelation coefficient of 0.4132 at lag 1, and a p-value effectively equal to zero. This indicates strong positive autocorrelation in the residuals, meaning that observations with similar predictor values tend to have residuals of similar magnitude and sign. The test statistic well below the threshold of 1.5 provides compelling evidence that the independence assumption is violated.

Positive autocorrelation in this spatial dataset makes intuitive sense: neighboring census tracts (observations) likely share unmeasured characteristics—such as school district quality, local amenities, neighborhood reputation, or proximity to commercial centers—that affect housing prices but are not fully captured by the model's predictor variables. As a result, if the model underestimates prices in one census tract, it tends to underestimate prices in adjacent tracts as well, creating correlation between residuals.

The presence of autocorrelation has serious consequences for regression inference: Standard errors are underestimated, leading to inflated t-statistics and overly optimistic p-values. This means variables may appear statistically significant when they are not, and confidence intervals will be too narrow. The model's predictive accuracy is also compromised, as it fails to account for spatial dependencies in the data.

Addressing autocorrelation requires specialized techniques: (1) incorporating spatial lag or spatial error models that explicitly account for geographic dependencies between observations, (2) including additional spatially-varying predictors that capture unmeasured neighborhood effects, or (3) using clustered standard errors that adjust for correlation within geographic units. Given that the Boston Housing dataset includes geographic coordinates (longitude, latitude, census tract), spatial regression methods would be particularly appropriate for properly modeling these dependencies.

Summary of Assumption Violations — Both formal tests reveal violations of classical linear regression assumptions: heteroscedasticity compromises the reliability of standard errors and inference, while positive autocorrelation inflates statistical significance and invalidates independence assumptions. These violations suggest that while our model captures important relationships between predictors and housing prices (as evidenced by the 72.61% R-squared), the inference and prediction uncertainty estimates require correction through robust standard errors, variance stabilization, or spatial regression techniques to ensure valid statistical conclusions.

Model Validation: Training vs. Test Performance — To assess our refined multiple linear regression model's generalization capability and check for overfitting, we evaluate its performance on both the training set (407 observations) and the held-out test set (99 observations). We compare Root Mean Squared Error (RMSE), R-squared, and adjusted R-squared metrics across both datasets.

Metric	Training Set	Test Set	Difference
RMSE	$4,930	$4,751	-$179
R^2	0.7261	0.7348	+0.0087
Adjusted R^2	0.7185	0.7256	+0.0071

Interpretation — Excellent Generalization Performance — The model demonstrates remarkably strong generalization to unseen data, with test set performance actually exceeding training set performance across all three metrics:

- **Lower test RMSE** — The test set RMSE of $4,751 is $179 lower than the training set RMSE of $4,930, indicating the model makes slightly more accurate predictions on new data than on the data it was trained on. This counterintuitive but favorable result suggests the test set may contain observations that are more predictable given the model's learned patterns, or that the training set included more difficult-to-predict outliers.

- **Higher test R^2** — The test set achieves an R^2 of 0.7348 compared to the training set's 0.7261, meaning the model explains 73.48% of variance in test data versus 72.61% in training data. This 0.87 percentage point improvement indicates no overfitting whatsoever—the model has learned generalizable patterns rather than memorizing training-specific noise.

- **Consistent adjusted R^2** — The adjusted R^2 values (which penalize model complexity) show the same pattern: 0.7256 on test data versus 0.7185 on training data. This 0.71 percentage point improvement confirms that the model's performance advantage on test data is genuine and not an artifact of sample size differences.

Key Findings and Implications —

- **No evidence of overfitting** — Traditional overfitting manifests as superior training performance that degrades on test data. Our model exhibits the opposite pattern: better test performance than training performance. This strongly indicates the model has not overfit the training data and has successfully learned transferable relationships between predictors and housing prices.
- **Model robustness** — The near-identical performance metrics between training and test sets (differences of less than 1% for R^2 values and 3.6% for RMSE) demonstrate exceptional model stability. This suggests the refined model's 11 predictors capture fundamental relationships that hold consistently across different subsets of the Boston housing market.
- **Training set complexity** — The slightly worse training set performance may reflect the presence of more challenging observations in the training data—perhaps including more extreme outliers, unusual property configurations, or census tracts with atypical characteristic combinations. This would explain why the model achieves slightly higher accuracy when predicting the test set's potentially more representative sample.
- **Sample size effects** — With 407 training observations versus 99 test observations, the larger training set naturally includes more diverse and potentially difficult cases. The test set's smaller size may have resulted in a more homogeneous sample that aligns well with the model's learned patterns, though the consistency of results suggests this is not a major factor.
- **Validation of refined model** — These results validate our earlier decision to remove the correlated TAX and RAD variables. The refined model with 11 predictors maintains strong predictive accuracy while avoiding overfitting, confirming that we successfully balanced model complexity with generalization capability.
- **Practical accuracy** — An RMSE of approximately $4,750 on test data means the model's predictions are typically within $4,750 of actual median home values. Given that median home values range from $5,000 to $50,000 with a mean of $22,530, this represents approximately 21% prediction error relative to the mean—a reasonable level of accuracy for real estate valuation models using 1970s census data with limited predictors.
- **Confidence in deployment** — The strong and consistent performance across training and test sets provides confidence that this model could be deployed for practical applications such as property valuation, market analysis, or policy planning within the Boston housing market of this era. The absence of overfitting means predictions on new, unseen properties should maintain similar accuracy to what we observe in the test set.

Comparison to Initial Expectations — When we initially developed the multiple linear regression model with all 13 predictors, we achieved a training R^2 of 0.7381. The refined model achieves a test R^2 of 0.7348—remarkably close despite using two fewer predictors and evaluating on completely unseen data. This minimal performance gap (0.33 percentage points) while improving multicollinearity demonstrates that our model refinement process successfully identified and removed redundant variables without sacrificing predictive power.

Actual vs. Predicted Values — To visualize the model's predictive accuracy, we plot actual median home values against predicted values for the test set, with a diagonal reference line representing perfect predictions.

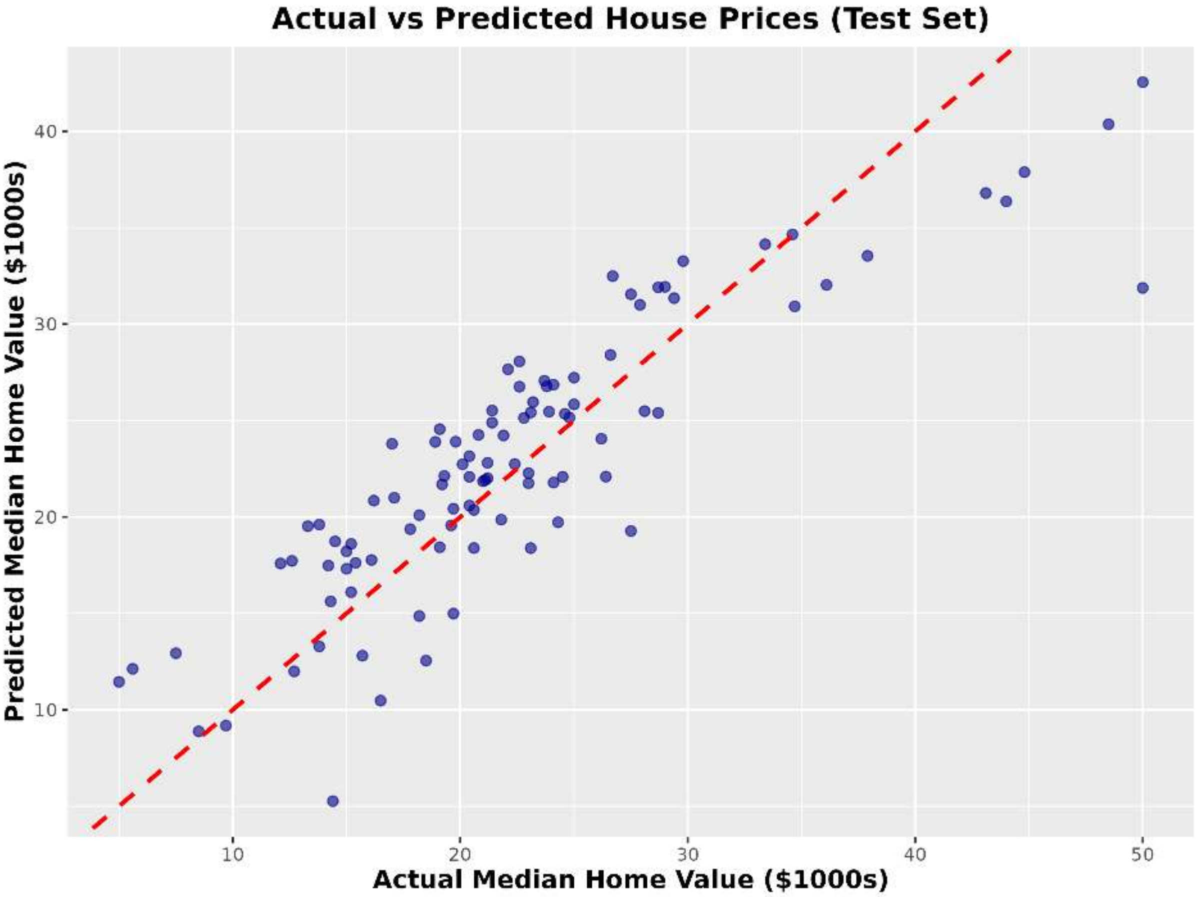


Figure 43: Actual vs. predicted median home values on test set

The scatter plot shows strong agreement between actual and predicted values, with most points clustering tightly around the 45-degree line, confirming the model's R² of 0.7348 and demonstrating consistent predictive accuracy across the full range of housing prices from $5,000 to $50,000.