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.

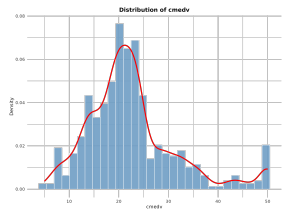

Figure 1: Distribution of CMEDV

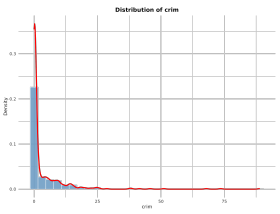

Figure 2: Distribution of CRIM

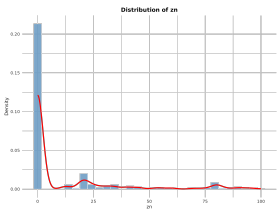

Figure 3: Distribution of ZN

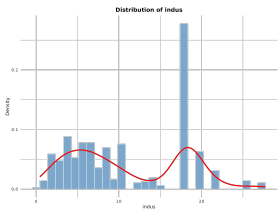

Figure 4: Distribution of INDUS

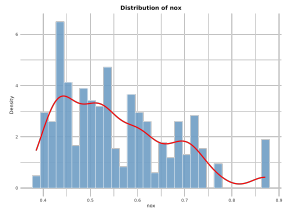

Figure 5: Distribution of NOX

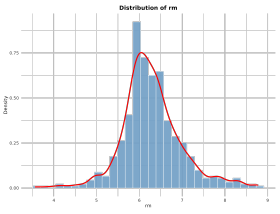

Figure 6: Distribution of RM

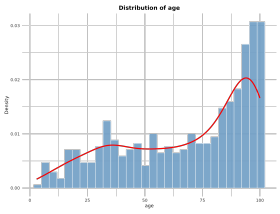

Figure 7: Distribution of AGE

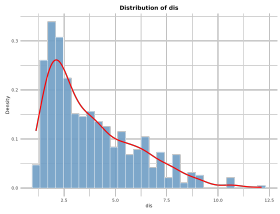

Figure 8: Distribution of DIS

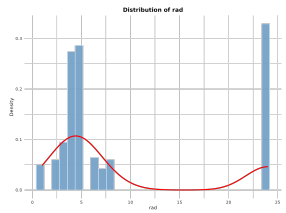

Figure 9: Distribution of RAD

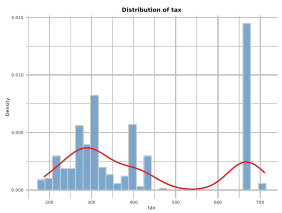

Figure 10: Distribution of TAX

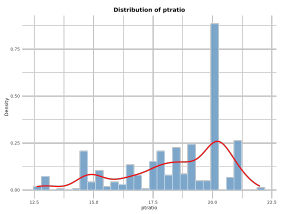

Figure 11: Distribution of PTRATIO

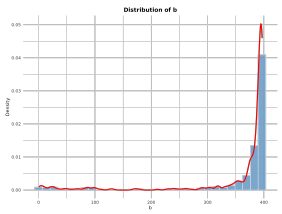

Figure 12: Distribution of B

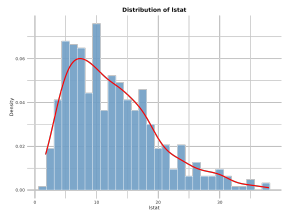

Figure 13: Distribution of LSTAT

Visualization

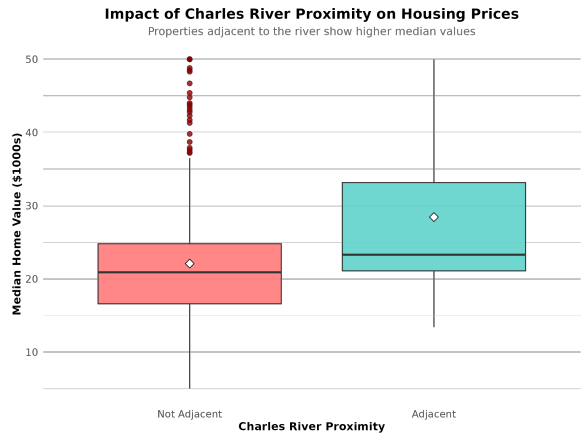

Figure 14: Properties by Charles River proximity

Description

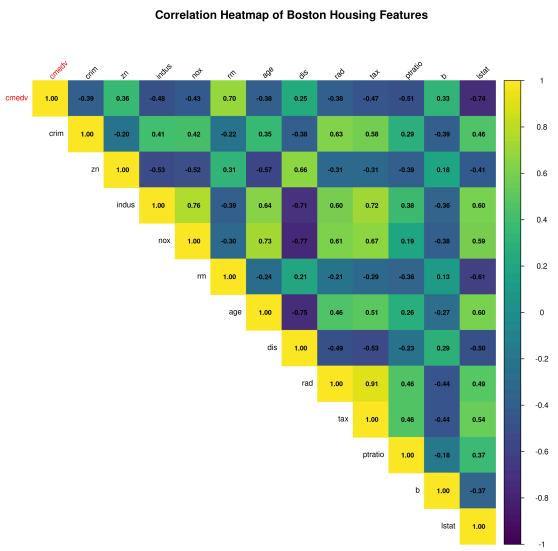

Figure 15: Correlation heatmap

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.

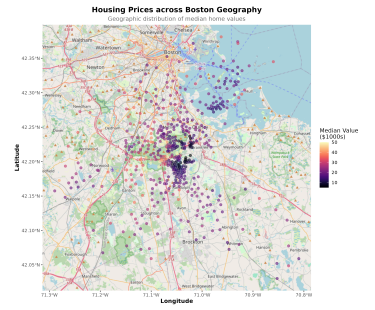

Figure 17: Geographic distribution of median home values

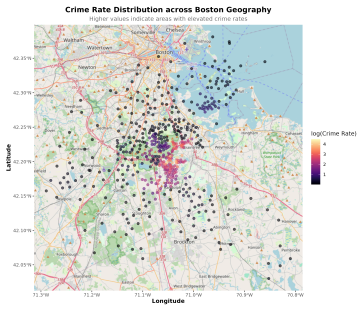

Figure 18: Geographic distribution of crime rate

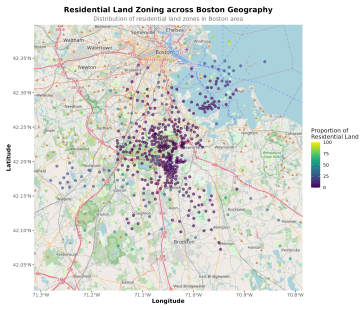

Figure 19: Geographic distribution of residential zoning

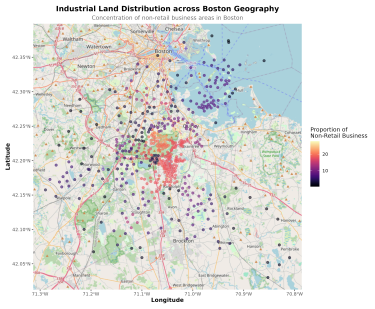

Figure 20: Geographic distribution of industrial areas

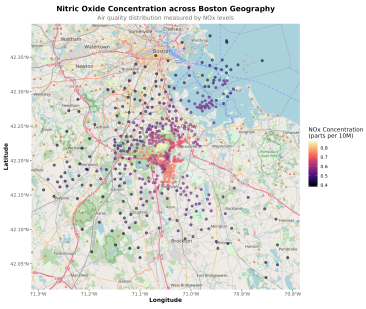

Figure 21: Geographic distribution of air pollution

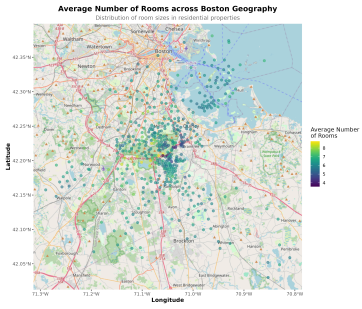

Figure 22: Geographic distribution of rooms per dwelling

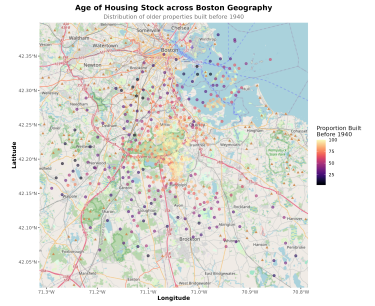

Figure 23: Geographic distribution of building age

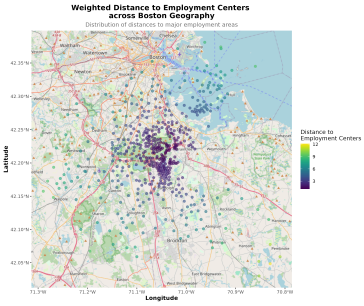

Figure 24: Geographic distribution of employment center distance

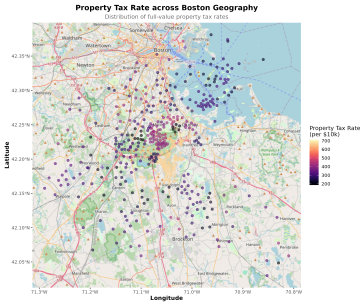

Figure 25: Geographic distribution of property tax rate

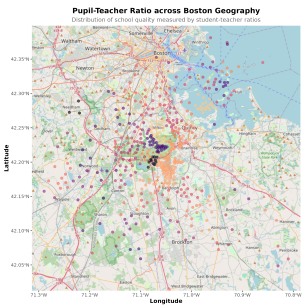

Figure 26: Geographic distribution of pupil-teacher ratio

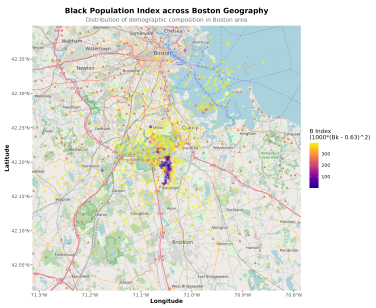

Figure 27: Geographic distribution of demographic variable

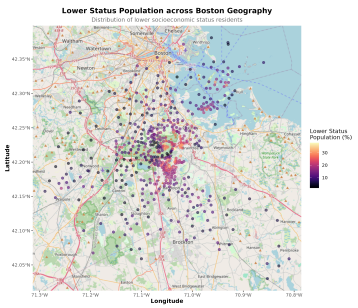

Figure 28: Geographic distribution of lower status population

Scatter Plot Analysis

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

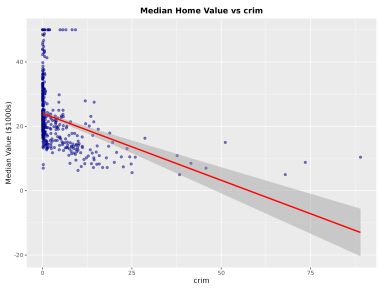

Figure 29: CMEDV vs CRIM

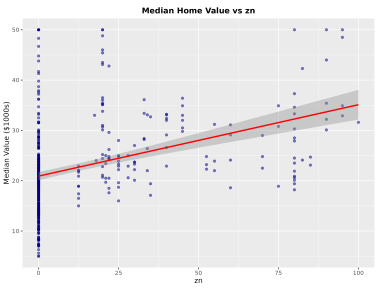

Figure 30: CMEDV vs ZN

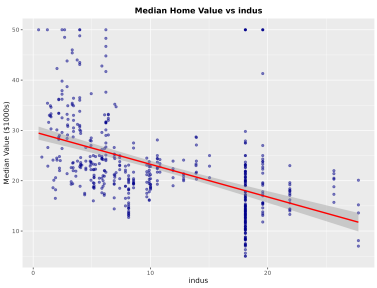

Figure 31: CMEDV vs INDUS

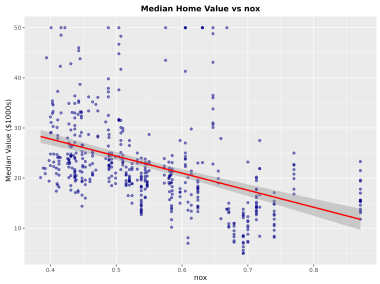

Figure 32: CMEDV vs NOX

Figure 33: CMEDV vs RM

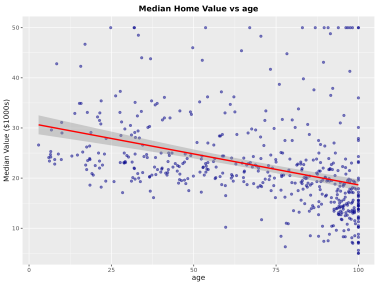

Figure 34: CMEDV vs AGE

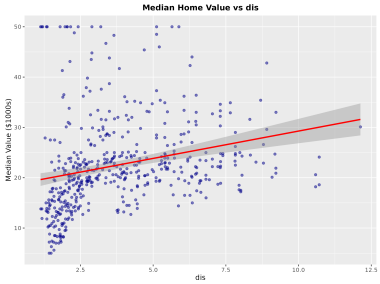

Figure 35: CMEDV vs DIS

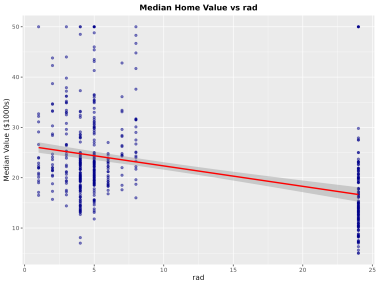

Figure 36: CMEDV vs RAD

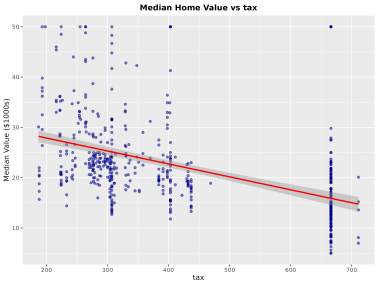

Figure 37: CMEDV vs TAX

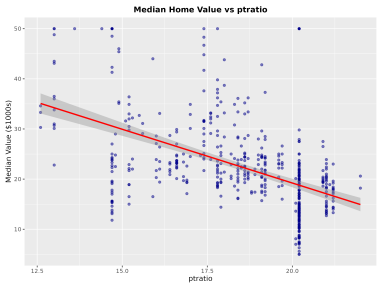

Figure 38: CMEDV vs PTRATIO

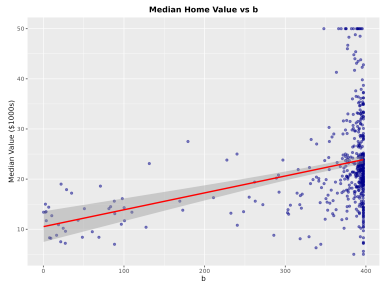

Figure 39: CMEDV vs B

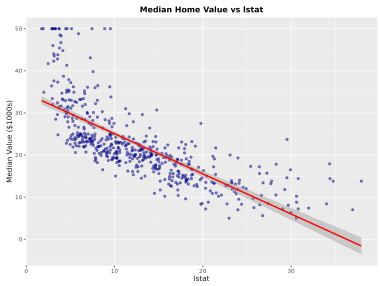

Figure 40: CMEDV vs LSTAT

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** ($0.7 \leq W < 0.9$) 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** ($W \geq 0.9$) 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.