BOSTON HOUSE PRICE PREDICTION

ELECTIVE PROJECT
MACHINE LEARNING

Dr. Uzma Abdullah
6.Oct, 2025

EXECUTIVE SUMMARY

This project examines the Boston Housing dataset to classify the key determinants of housing prices and develop a reliable consistent predictive model. The dataset consists of 506 observations with 13 variables, all numeric and complete, with no missing value, covering socioeconomic, structural, and environmental factors influencing median home values (MEDV). Exploratory data analysis revealed skewness in numerous features (e.g., crime rate, socioeconomic status, and taxes) and a capped distribution for housing prices at $50,000. Strong correlations were observed between MEDV and variables such as number of rooms (RM, positive) and percentage of lower-status population (LSTAT, negative), education quality (PTRATIO), pollution levels (NOX), crime rate (CRIM), and proximity to employment centers (DIS) also demonstrated significant impacts. To address skewness and improve model assumptions, a log transformation was applied to MEDV, and multicollinearity checks led to the removal of highly correlated variables like TAX. The final absolute multiple linear regression model achieved an R^2 of approximately 0.77 on both training and testing data, with a low RMSE, indicating good explanatory power and predictive accuracy and precision. Outcomes endorse that socioeconomic conditions, environmental excellence, and structural attributes and characteristics are the utmost critical drivers of housing prices in Boston, with larger homes, neat cleaner environments, better graded schools, and desirable appropriate locations commanding higher values, whilst disadvantaged neighbourhoods face depressed prices.

BUSINESS PROBLEM OVERVIEW

The real estate market is heavily influenced by multiple economic, social, and environmental factors, making accurate projection of housing prices acute for shareholders such as real estate developers, investors, policymakers, and homebuyers. In cities like Boston, housing affordability and valuation are mainly imperative given the range of neighbourhoods and the extensive range of socioeconomic conditions. However, traditional out-dated valuation methods often overlook the combined effect of factors like crime rates, access to amenities, environmental quality, education standards, and structural attributes of homes. This project aims to address the business challenge of identifying the key drivers of housing prices and developing a predictive model that can estimate property values with high accuracy. By doing so, stakeholders can make data-driven decisions: developers can better target investments, financial institutions can refine mortgage lending strategies, policymakers can design housing interventions, and homebuyers can make more informed purchasing choices. Eventually, solving this problem helps reduce ambiguity in real estate estimation, expands market efficiency and productivity, and specifies insights into the collective and conservational factors shaping housing demand in urban areas like Boston.

PROBLEM STATEMENT:

The challenge is to accurately predict housing prices in the Boston area by analysing the impact of socioeconomic, structural, and environmental factors. The goal is to build a reliable regression model that identifies the key drivers of home values and provides actionable insights for real estate stakeholders, policymakers, and homebuyers.

DATA OVERVIEW

The dataset used in this project is the Boston Housing dataset, which contains 506 observations and 13 variables describing various characteristics of residential areas in Boston. All variables are numeric, with no missing or duplicate values, making the dataset clean and suitable for statistical analysis and modelling.
Target Variable: MEDV – Median value of owner-occupied homes (in $1000s).
Independent Variables:

- CRIM – Per capita crime rate by town.
- ZN – Proportion of residential land zoned for lots over 25,000 sq. ft.
- INDUS – Proportion of non-retail business acres per town.
- CHAS – Charles River dummy variable (1 if tract bounds river; 0 otherwise).
- NOX – Nitric oxide concentration (parts per 10 million).
- RM – Average number of rooms per dwelling.
- AGE – Proportion of owner-occupied units built prior to 1940.
- DIS – Weighted distance to five Boston employment centers.
- RAD – Index of accessibility to radial highways.
- TAX – Property tax rate per $10,000.
- PTRATIO – Pupil–teacher ratio by town.
- LSTAT – Percentage of lower-status population.
- Key Characteristics of the Data:
- The dataset covers socioeconomic, environmental, and structural factors.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV	
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV	
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	5.33	36.2

OBSERVATIONS FROM DATA OVERVIEW

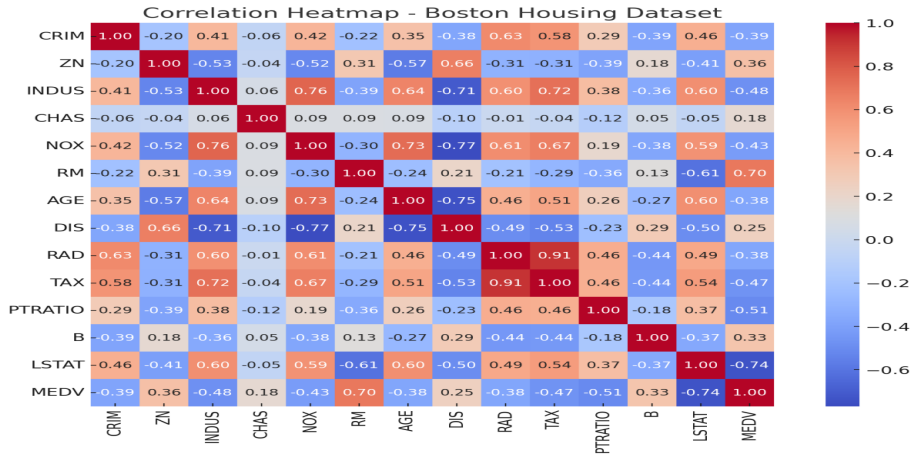
Shape of the dataset, the dataset contains 506 rows and 13 columns. Each row represents a housing tract in the Boston area. All 13 variables are numeric (float64 or int64). No missing values are present (all columns have 506 non-null entries). Features include socioeconomic (LSTAT), structural (RM, AGE), and geographical variables (DIS, RAD, TAX), plus the target variable MEDV. The variable MEDV is the target variable, representing the Median value of homes (in $1000s). Statistical summary shows, CRIM (crime rate) is highly skewed: minimum is close to 0, but maximum is very high. RM (average number of rooms) ranges from 3.56 to 8.78. LSTAT (% lower status population) varies between 1.7% and 37%, showing diversity in socioeconomic conditions. MEDV (house price) ranges from 5.0 to 50.0 (capped at 50, suggesting some values hit a maximum).

Range Index: 506 entries, 0 to 505		
Data columns (total 13 columns):		
# Column	Non-Null Count	Dtype
--- ------ -------------- -----		
0 CRIM	506 non-null	float64
1 ZN	506 non-null	float64
2 INDUS	506 non-null	float64
3 CHAS	506 non-null	int64
4 NOX	506 non-null	float64
5 RM	506 non-null	float64
6 AGE	506 non-null	float64
7 DIS	506 non-null	float64
8 RAD	506 non-null	int64
9 TAX	506 non-null	int64
10 PTRATIO	506 non-null	float64
11 LSTAT	506 non-null	float64
12 MEDV	506 non-null	float64
dtypes: float64(10), int64(3)		
memory usage: 51.5		

BOSTON HOUSING DATASET - SUMMARY STATISTICS

Feature	count	mean	std	min	25%	50%	75%	max
CRIM	506.0	3.614	8.602	0.006	0.082	0.257	3.677	88.976
ZN	506.0	11.364	23.322	0.0	0.0	0.0	12.5	100.0
INDUS	506.0	11.137	6.86	0.46	5.19	9.69	18.1	27.74
CHAS	506.0	0.069	0.254	0.0	0.0	0.0	0.0	1.0
NOX	506.0	0.555	0.116	0.385	0.449	0.538	0.624	0.871
RM	506.0	6.285	0.703	3.561	5.885	6.208	6.624	8.78
AGE	506.0	68.575	28.149	2.9	45.025	77.5	94.075	100.0
DIS	506.0	3.795	2.106	1.13	2.1	3.207	5.188	12.126
RAD	506.0	9.549	8.707	1.0	4.0	5.0	24.0	24.0
TAX	506.0	408.237	168.537	187.0	279.0	330.0	666.0	711.0
PTRATIO	506.0	18.456	2.165	12.6	17.4	19.05	20.2	22.0

B	506.0	356.674	91.295	0.32	375.378	391.44	396.225	396.9
LSTAT	506.0	12.653	7.141	1.73	6.95	11.36	16.955	37.97
MEDV	506.0	22.533	9.197	5.0	17.025	21.2	25.0	50.0



The correlation heatmap highlights the relationships between predictors and housing prices (MEDV). The strongest positive correlation is observed between RM (average number of rooms per dwelling) and MEDV (~+0.70), indicating that larger homes are more valuable. Conversely, the strongest negative correlation is with LSTAT (percentage of lower-status population, ~–0.74), showing that socioeconomic disadvantage reduces house prices. Other important negative relationships include PTRATIO (–0.51), NOX (–0.43), TAX (–0.47), and INDUS (–0.48), reflecting the impact of education quality, pollution, taxes, and industrialization. Positive but weaker correlations exist with ZN (zoned land, +0.36) and CHAS (proximity to Charles River, +0.18). Additionally, very high correlations among independent variables (e.g., RAD and TAX, ~+0.91) suggest multicollinearity risks that must be addressed in regression modelling.

STATISTICAL SUMMARY OBSERVATIONS:

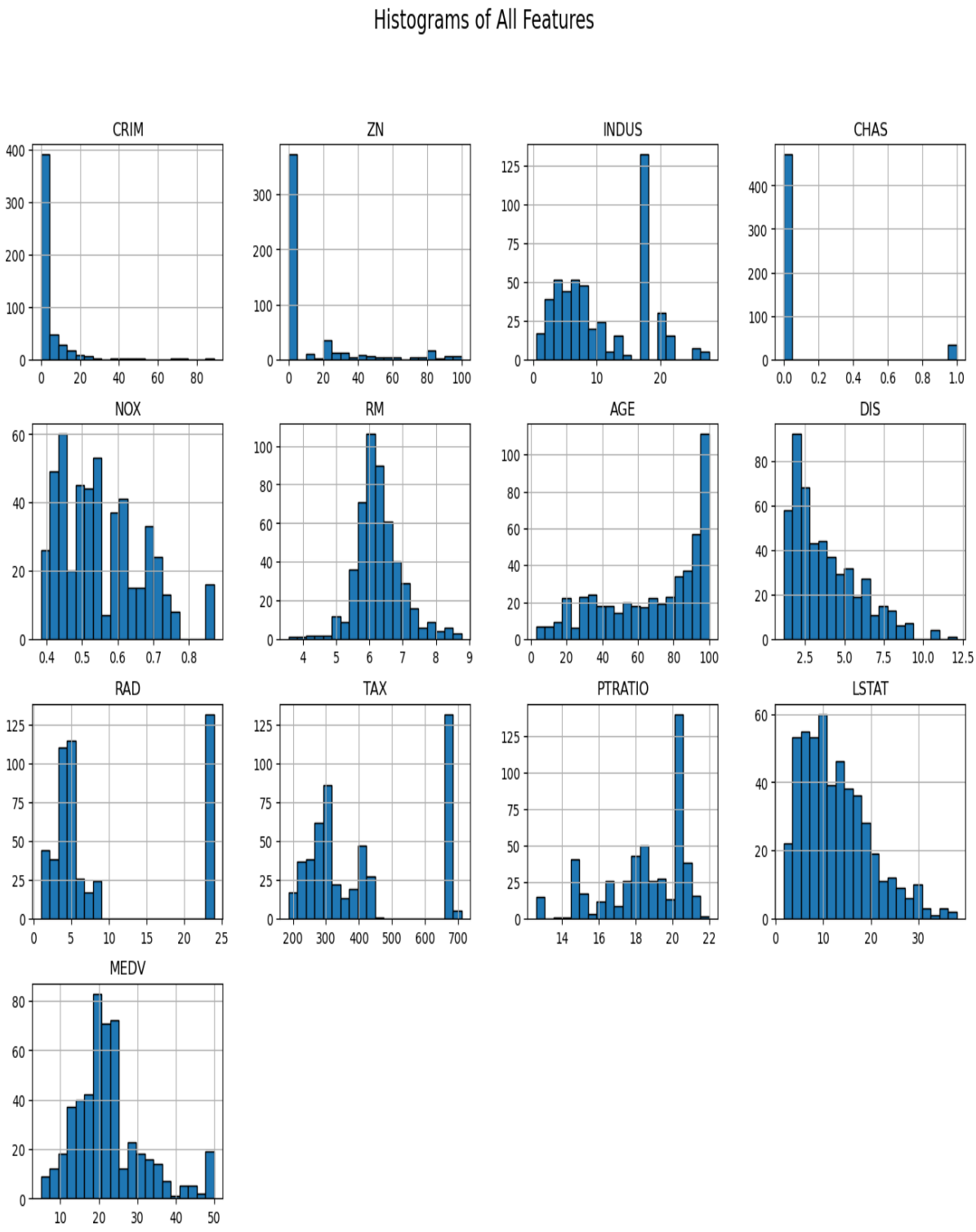
In general, all variables are numeric. Statistics shows range (min, max), central tendency (mean, median approx. via 50%), and spread (std). The key predictors are CRIM (Crime rate): Min = 0.006, Max ≈ 88.97, very skewed (most areas low crime, some very high). ZN (Zoned land): Many zeros, but max = 100, indicates some tracts are fully zoned for large lots. INDUS (Industrial proportion): Range 0.46 – 27.74, shows wide variation in industrial zones. NOX (Nitric oxide concentration): Mean ≈ 0.55, range 0.38 – 0.87.RM (Average rooms): Mean ≈ 6.28, range 3.56 – 8.78, most houses have 5–7 rooms. AGE (Older homes): Mean ≈ 68, with some tracts nearly 100% older homes. DIS (Distance to employment centers): Mean ≈ 3.79, range 1.1 – 12.1. TAX (Property tax): Range 187 –711 per $10,000. PTRATIO (Pupil–teacher ratio): Range 12.6 – 22.0. LSTAT (% lower status): Mean ≈ 12.7, range 1.7 – 37.9. Target variable (MEDV – Median Home Value) Mean ≈ 22.5 ($22,500). Min = 5.0, Max = 50.0 clearly capped at 50 (price ceiling in dataset). Standard deviation ≈ 9.2, showing a wide spread in home price

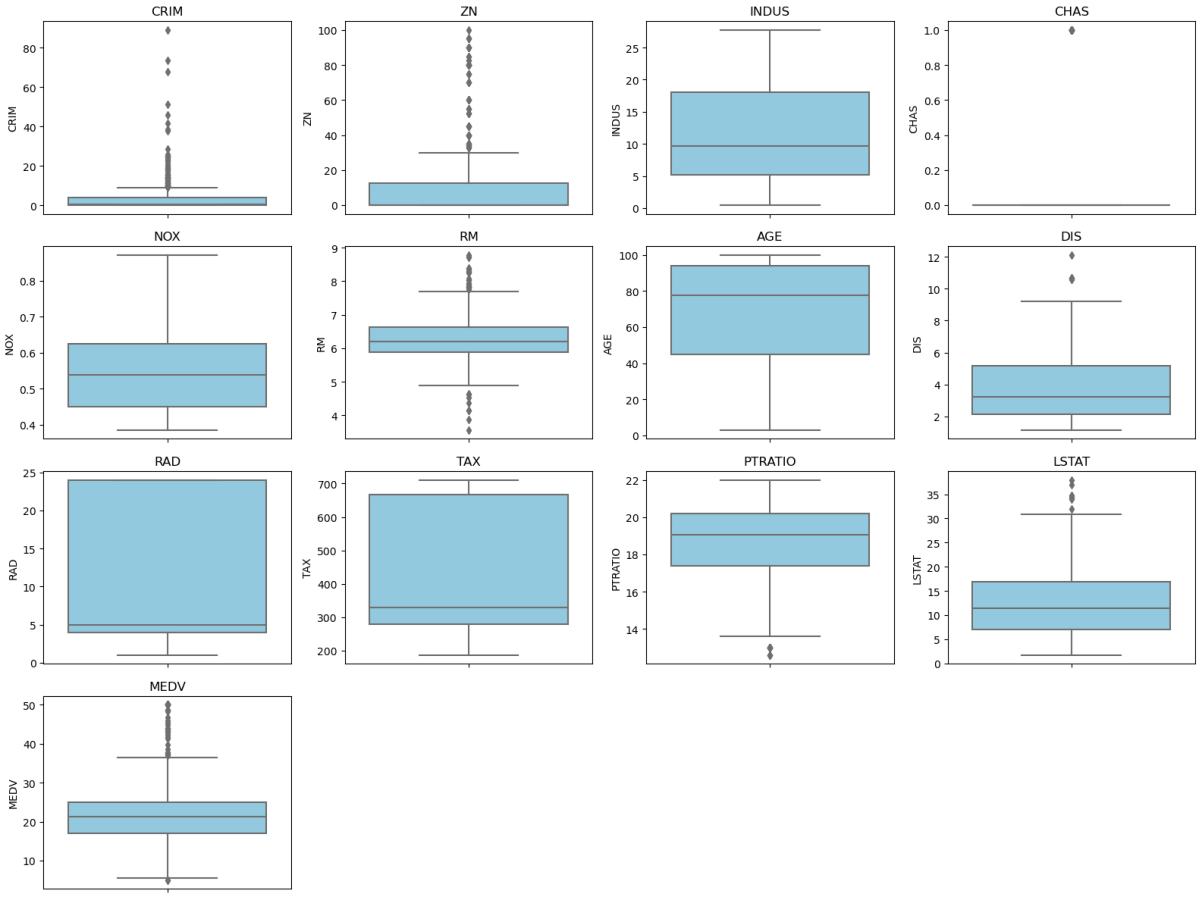
The dataset shows skewed distributions in crime rate, zoning, and housing prices. The cap at 50 for MEDV must be noted, as it may affect regression analysis. The Boston Housing dataset (506 rows × 13 columns) usually has no duplicate rows. As duplicates prints 0, so no duplicate rows found in the dataset. All 506 entries are unique. The summary statistics give you a clear picture of the distribution, spread, and range of each variable.

CODE FOR SUMMARY STATISTICS:

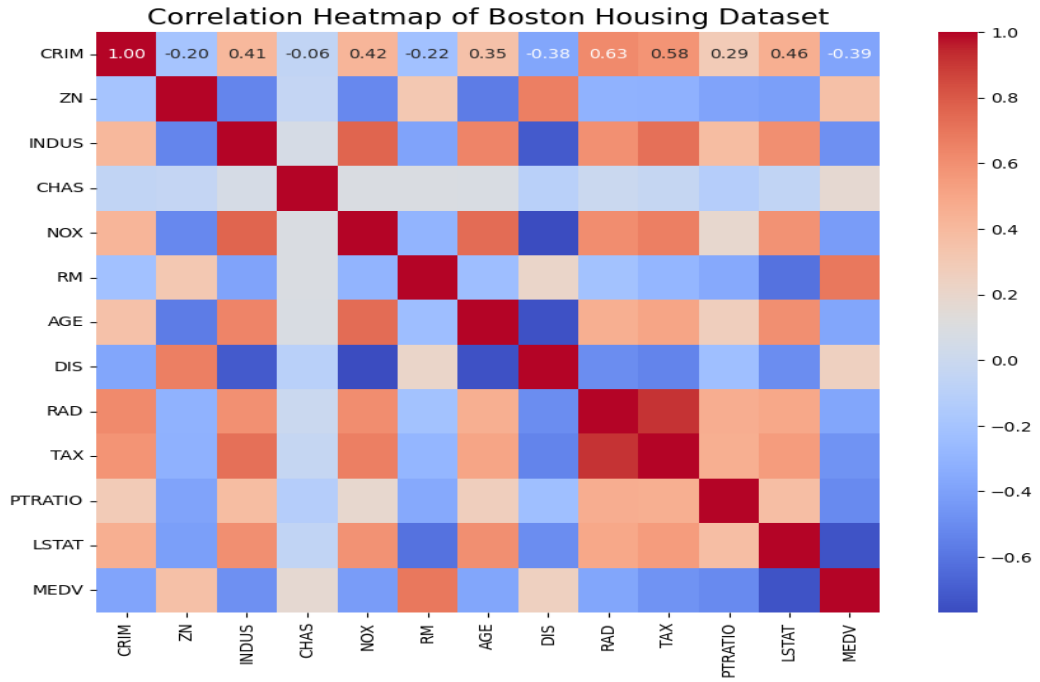
Shape of dataset: 506 rows × 13 columns. All variables are numeric (float64 or int64), no categorical columns. CRIM (crime rate): Highly skewed, min ≈ 0.006, max ≈ 88.9. • RM (avg rooms): Mean ≈ 6.28, range 3.56 – 8.78 (most homes ~5–7 rooms). • AGE (older homes): Mean ≈ 68%, many tracts have older houses. • LSTAT (% lower status): Mean ≈ 12.7, range 1.7 – 37.9 → large variation in socio-economic status. • MEDV (house prices, target): Mean ≈ 22.5, min = 5, max = 50 → capped at 50 (price ceiling). No missing values detected

VISUAL SUMMARY STATISTICS





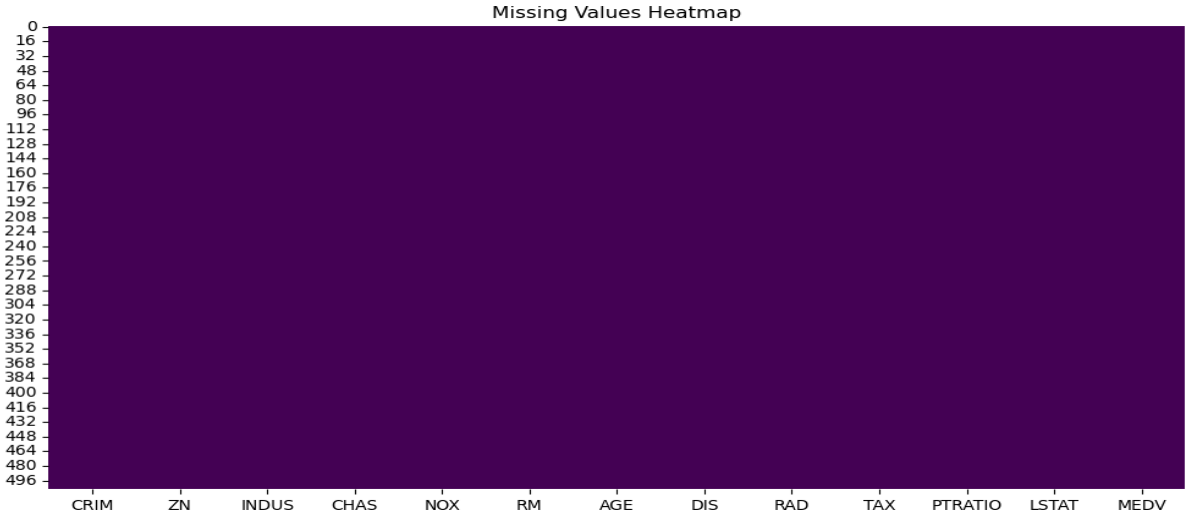
Most features are not normally distributed e.g. CRIM (crime rate) is heavily right-skewed. RM (average rooms) and MEDV (home values) look closer to normal but still have some skew. Outliers are visible in CRIM, ZN, B, and MEDV. MEDV is capped at 50, meaning high-value homes are truncated. Boxplots confirm extreme values in crime rate, taxes, and pupil–teacher ratio.



Observations from the Heatmap • Strong positive correlation with MEDV (house price): RM (avg rooms, ~+0.70) more rooms thus higher prices. • B (proportion of Black population, ~+0.33). Strong negative correlation with MEDV: • LSTAT (% lower status, ~–0.74), higher % of lower-status population, lower prices. PTRATIO (pupil–teacher ratio, ~–0.51), higher ratio, lower prices. NOX (pollution, ~–0.43) and DIS (distance to employment centres, ~–0.38). • Highly correlated predictors (multicollinearity risk): TAX and RAD (~+0.91). DIS and NOX (~–0.77). CRIM and RAD (~+0.63). These correlations suggest that LSTAT and RM are the most important predictors of house prices.

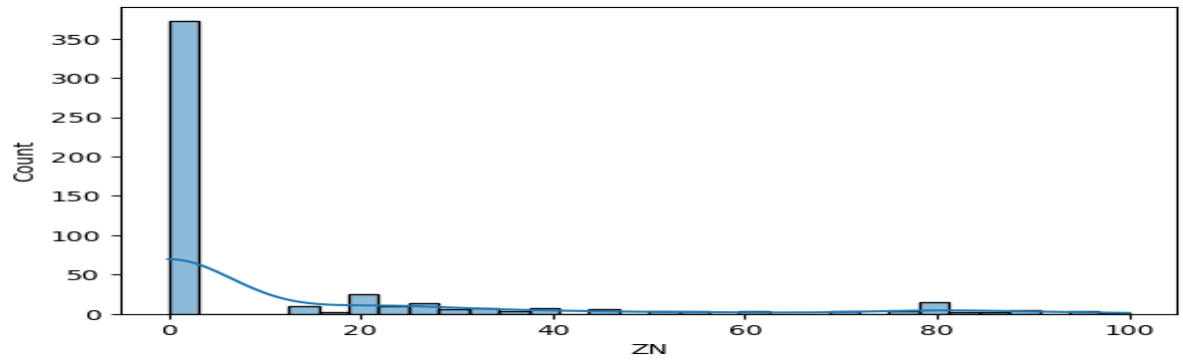
EXPLORATORY DATA ANALYSIS AND DATA PREPROCESSING

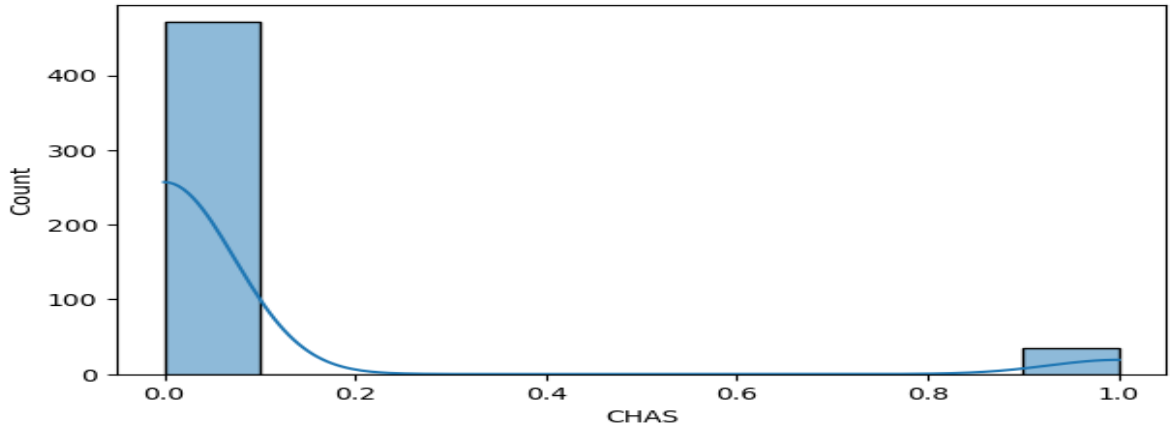
Expected Observation for Boston Housing: Typically, the Boston Housing dataset has no missing values, the output should be all zeros. If you see non-zero counts, you'll need to handle them (drop rows or impute with mean/median). After dropping rows -> Shape: (506, 13) After filling missing values -> Shape: (506, 13)

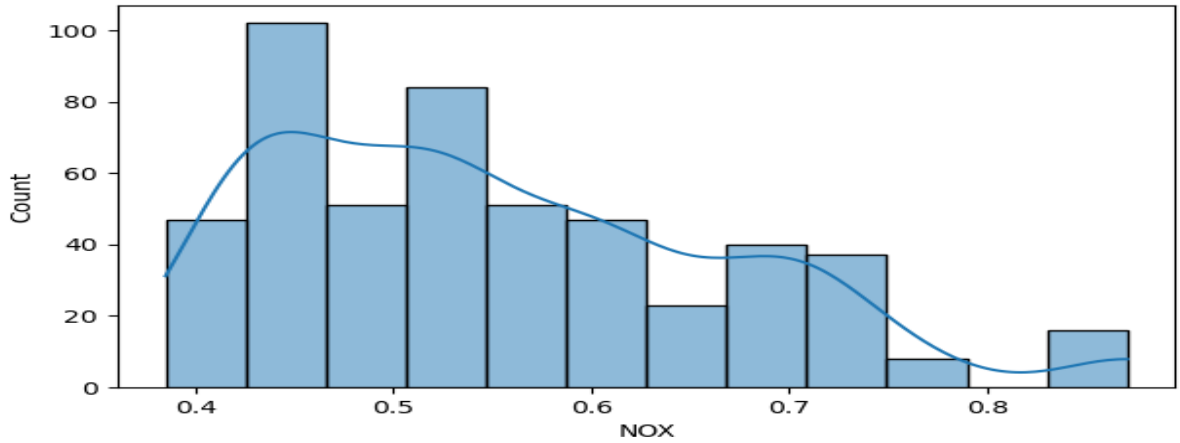


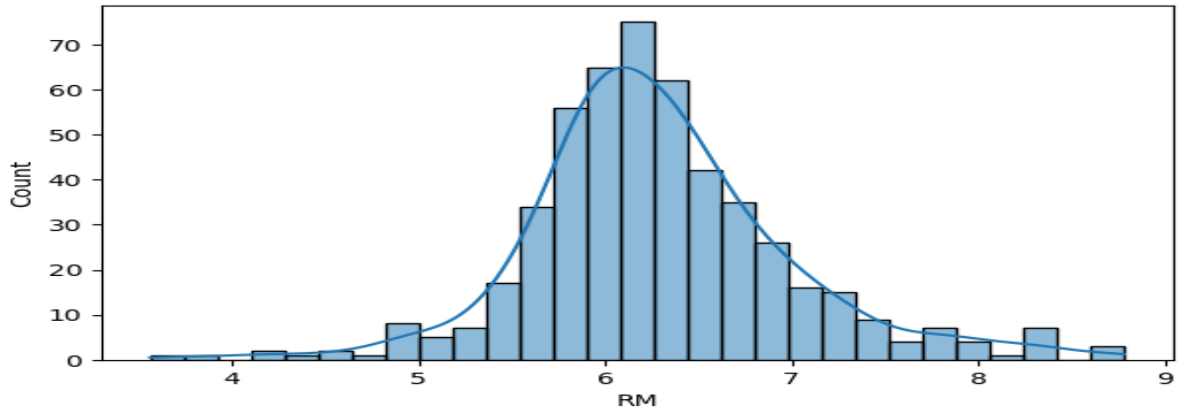
Expected Observation for Boston Housing: Typically, the Boston Housing dataset has no missing values, output should be all zeros. If we see non-zero counts, we will need to handle them (drop rows or impute with mean/median). Expected Observation (Boston Housing dataset): The heatmap should be completely filled (no yellow/white lines), meaning no missing values. If there are vertical lines, it means that column has missing entries.

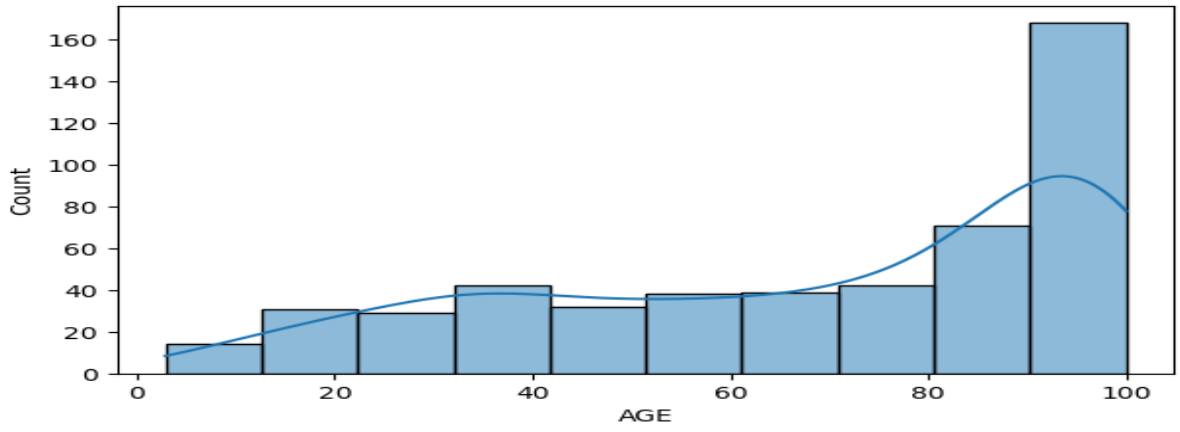
DISTRIBUTION OF COLUMNS:

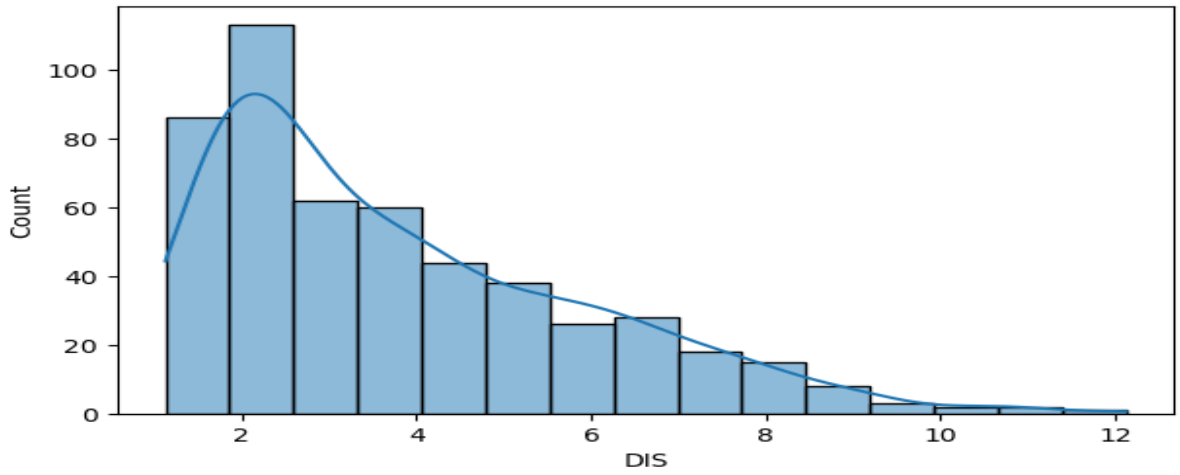


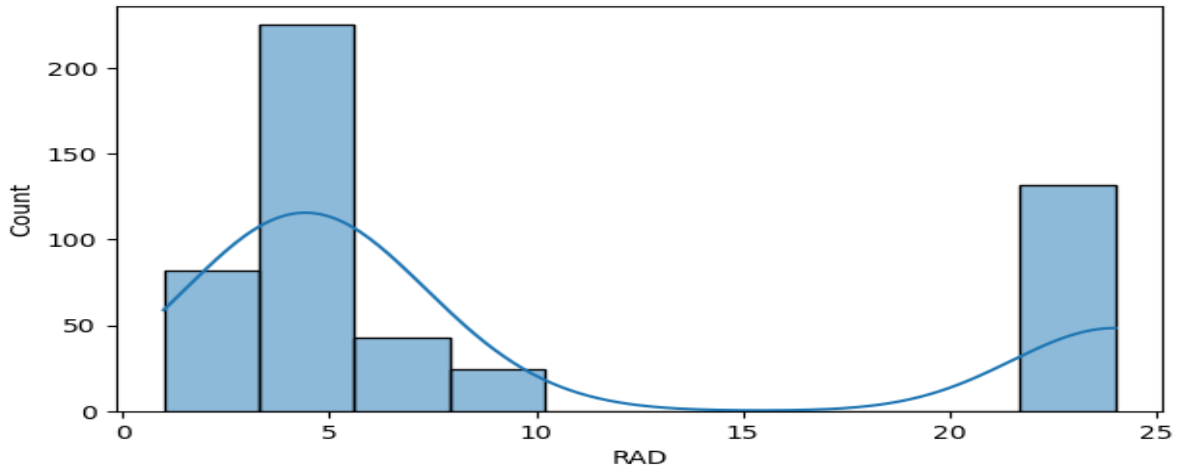

Count
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
CHAS

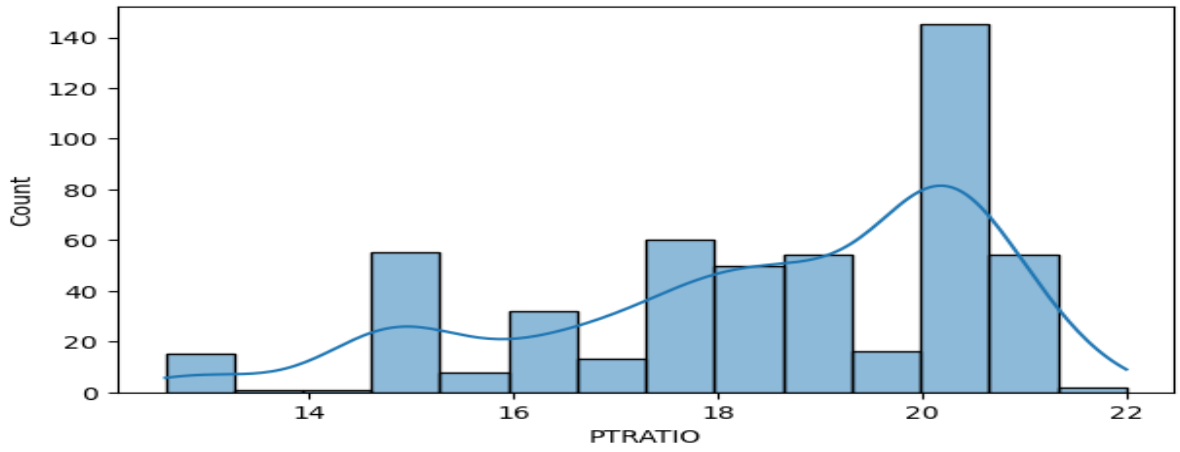

Count
100
80
60
40
20
0
0.4
0.5
0.6
0.7
0.8
NOX

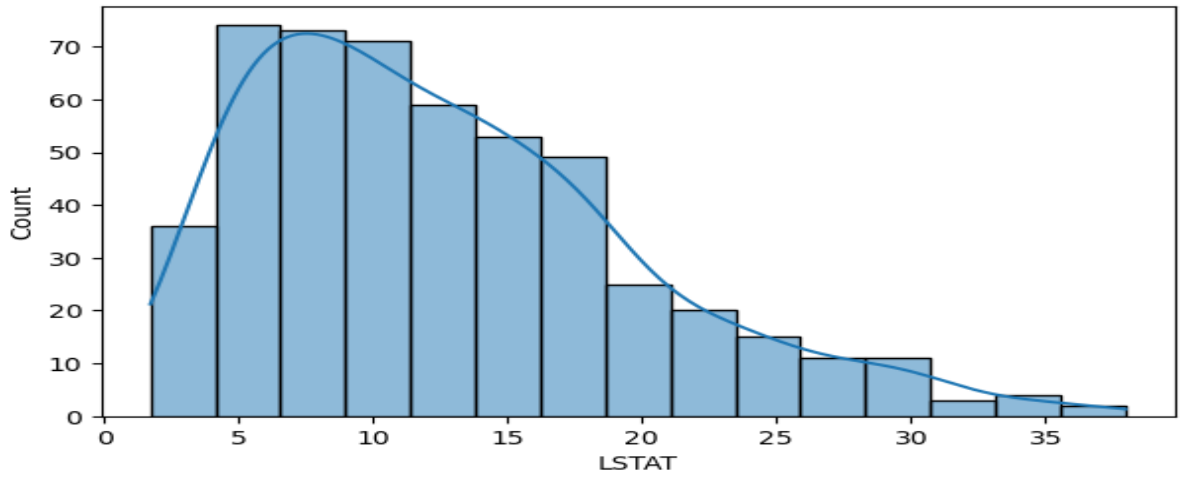

Count
70
60
50
40
30
20
10
0
4
5
6
7
8
9
RM

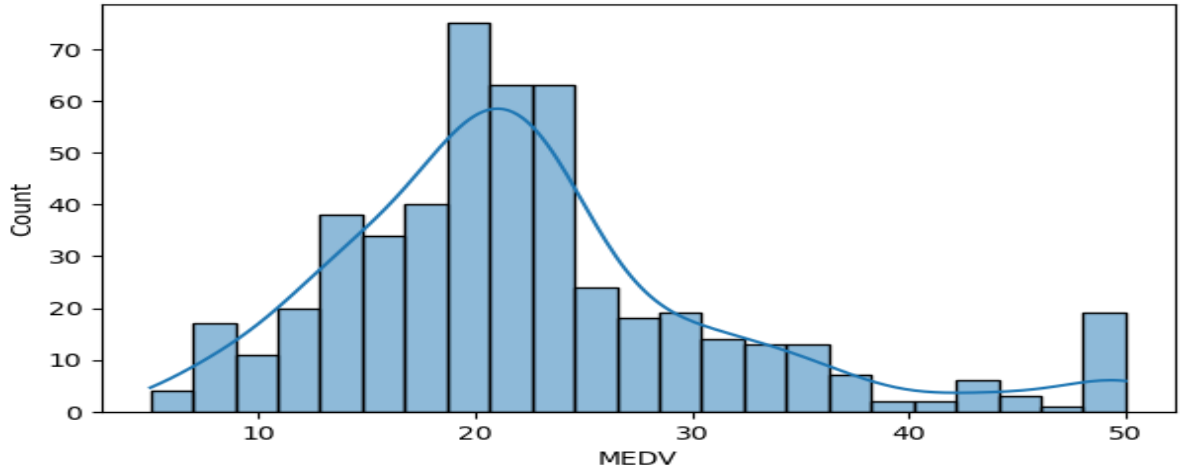

Count
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
AGE


Count
100
80
60
40
20
0
2
4
6
8
10
12
DIS


Count
200
150
100
50
0
0
5
10
15
20
25
RAD


Count
140
120
100
80
60
40
20
0
14
16
18
20
22
PTRATIO





DESCRIPTION OF FEATURES:

1. **CRIM (Per capita crime rate)**

The Histogram shows strong right skew position. Most neighbourhoods' communities have very low crime rates, but a few extreme outliers have very high crime. Boxplot shows outliers clearly visible because Skewness indicates we may need log transformation before modelling.

2. **ZN (% Residential land zoned for lots which are greater than 25,000 sq ft)**

The histogram shows mostly 0 having few neighbourhoods having high zoning. The boxplot reflects long tail to the right because this is a sparse variable as most areas don't have large-lot zoning, so it may not contribute much to prediction.

3. **INDUS (% non-retail business acres per town)**

The histogram has two clusters, one around 5–10, another around 18–20. The boxplot reflects some variation but fewer extreme outliers. This is because it helps to measure industrialization of area which might negatively correlate with house prices.

4. **CHAS (Charles River dummy variable)**

Histogram: Only 0 or 1 (binary) with the boxplot doesn't seems to be meaningful here (only two values) which Indicates proximity to Charles River, can directly impact housing values.

5. **NOX (Nitric Oxide concentration)**

The histogram appears to be slight right skew, concentrated around 0.45–0.6 and the boxplot reflects the outliers at higher NOX levels. The possible reason could be Pollution, thus higher NOX likely reduces housing prices.

6. **RM (Average rooms per dwelling)**

In this the histogram is about nearly normal, mean nearly equal to 6.3, with some outliers seems to be greater than 8. The boxplot shows the few high outliers. It is one of the strongest predictors showing more rooms leads to higher house prices.

7. **AGE (% owner-occupied units built before 1940)**

The Histogram is skewed left, many older houses and the boxplot shows the outliers at very low ages. The reason is it captures historical housing standard which may interact with other features like crime.

8. **DIS (Weighted distance to employment centres)**

The histogram is right skew showing many neighbourhoods are closer to jobs. The boxplot: Some high-distance outliers. Hence, distance affects accessibility and nearer homes frequently are more valuable.

9. **RAD (Index of accessibility to radial highways)**

In this feature the histogram is highly discrete depicting big spikes at 4 and 24. The boxplot outliers seems to be common. It is due to reason that the road access is categorical-like which may need dummy encoding.

10. **TAX (Property tax rate)**

The tax histogram illustrates the peaks at certain brackets (discrete isolated jumps) and the boxplot has large spread, with elevated outliers. It can be seen that the tax rate directly affects affordability and people interest.

11. **PTRATIO (Pupil–teacher ratio)**

Here the histogram looks Centred around 17–20 and the boxplot has relatively tight distribution. The argument behind is the Lower PTRATIO as there are better schools leads to higher prices.

12. **LSTAT (% lower status population)**

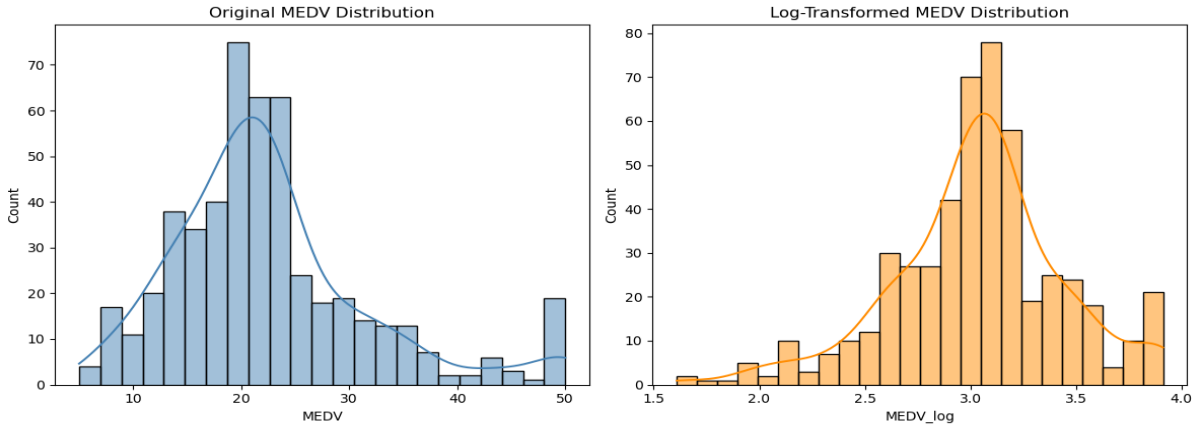
In this variable the histogram is about right skew with most values less than 15%. Some outliers at high values (30–40) in the boxplot which shows that the lower status position is the strong negative predictor of house price.

13. **MEDV (Median house value, $1000s)**

The histogram shows the Bell-shaped but capped at 50 (maximum artificial ceiling). The outliers at upper bound 50 in the boxplot, the target variable cap creates a modelling challenge (editing).

DEDUCTION OF UNIVARIATE ANALYSIS

The above analysis shows various variables are skewed (CRIM, ZN, RAD, LSTAT) which may need transformation and alteration. Some variables are discrete/categorical-like (CHAS, RAD, TAX) which might necessitate encryption or encoding. The target dependent variable, Median value of houses (MEDV) is capped by replacing extreme values or outliers with threshold so the models might underestimate very high-value houses. The Univariate (single-variable) analysis is the foremost step to indicate where to clean, transform, or plot features before constructing predictive models for the analysis. Since MEDV ($1000s) is the dependent variable, the univariate analysis shows it has slightly right-skewed distribution, capped at 50, hence can affect regression models as the linear regression assumes that the residuals are normally distributed. Further, in order to decrease the skewness and stabilize variance, we apply a logarithmic transformation to MEDV. This alteration reduces higher values and gives lower values, formulating the distribution closer to normal values.



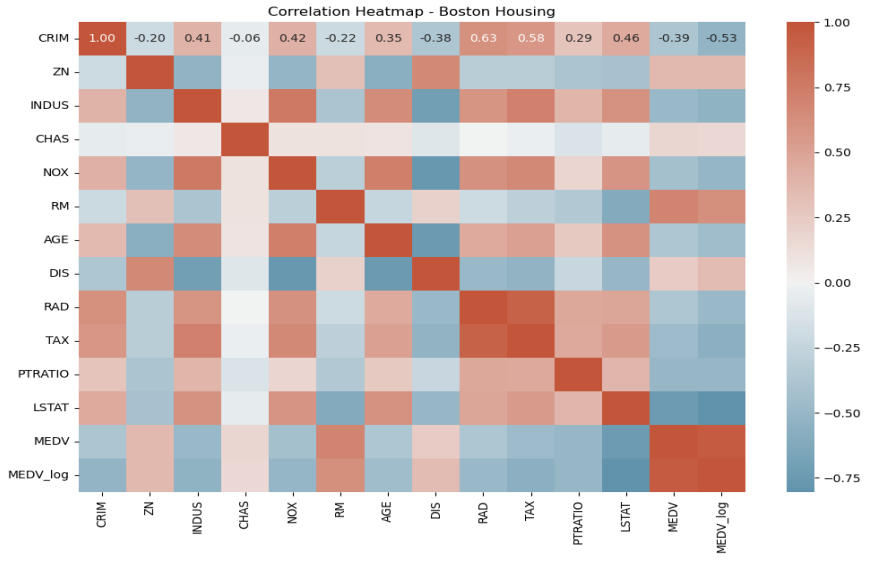
The dependent variable MEDV (median value of homes in $1000s) was found to be slightly right-skewed in its original form, with a visible cap at 50.0. This non-normality violates regression assumptions and can reduce model performance. To address this, we applied a logarithmic transformation. The original MEDV was nearly 1.1 (moderately right skewed) with compressed distribution at higher values (due to cap at 50) has changed to log transformed MEDV-log with skewness nearly equal to 0.2 with distribution much closer to normal. It shows the high values are "compacted" and low values are "stretched" constructing a smoother, bell-shaped spread. The purpose for this exercise is to:

- Firstly, it enhances the Regression Fit by training MEDV_log models which satisfies the assumption of normally distributed residuals.
- Secondly, it reduces heteroscedasticity as variance of errors becomes more stable through unusual ranges of housing prices.
- Thirdly, while using regression coefficients they can be interpreted in terms of percentage changes rather than raw dollar fluctuations. Here the coefficient of 0.05 on RM means that each additional room is associated with about 5% higher median house price while keeping other factors constant.

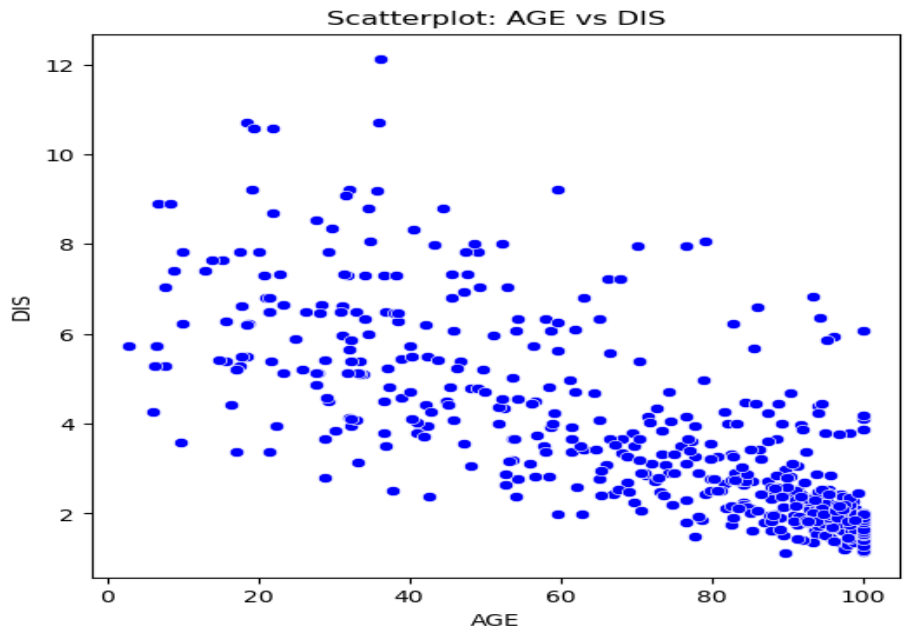
BIVARIATE ANALYSIS

Furthermore, before heading into regression we proceed to Bivariate analysis in order to see how one variable relationship changes with other (correlation heatmap & scatterplots RM/LSTAT vs MEDV) and which predictors are most strongly related to house prices we use:

1. Heatmap showing correlations between all variables.
2. Scatterplots specifically between significant predictors (RM, LSTAT, CRIM, TAX) and the target (MEDV or MEDV_log).

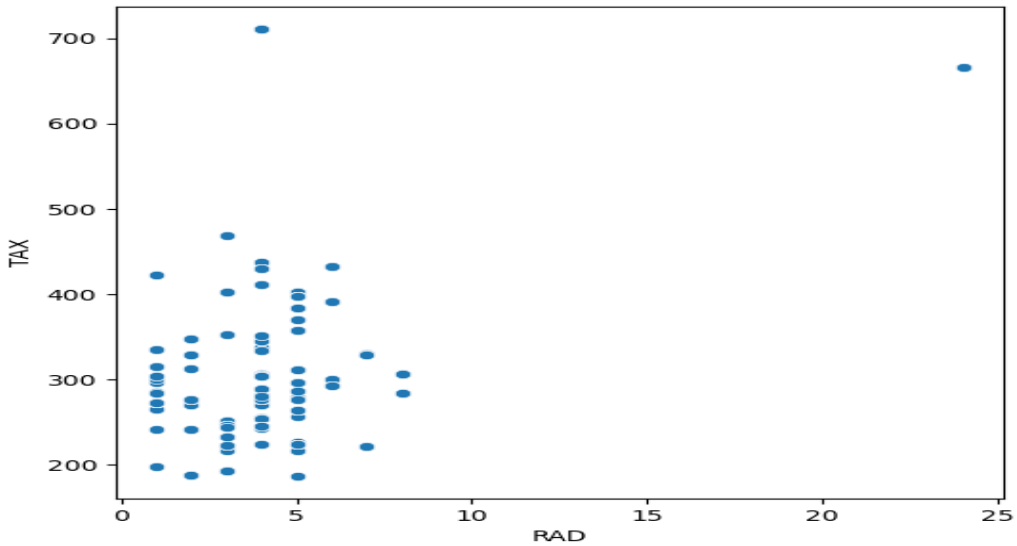


THE RELATIONSHIP BETWEEN AGE AND DIS



Strong negative correlation exists between AGE (older houses) and DIS (distance to employment centres). It is seen that older houses tend to be closer to the city centre, while newer houses are located farther away. Hence, the graph imitates that the development city centres in urban area have older housing standard. Another logic could be that the Boston employment centres are based in developed towns since before 1940 which are mostly occupied by owners.

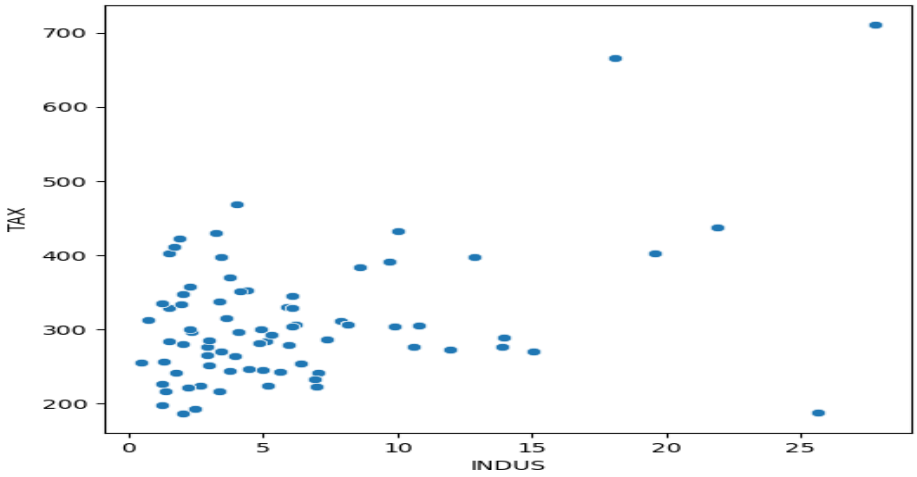
THE RELATIONSHIP BETWEEN RAD AND TAX



The above graph shows the correlation between RAD and TAX is very high, whereas no trend is visible between these two variables. Hence, the strong correlation might be due to outliers. However, the correlation between TAX

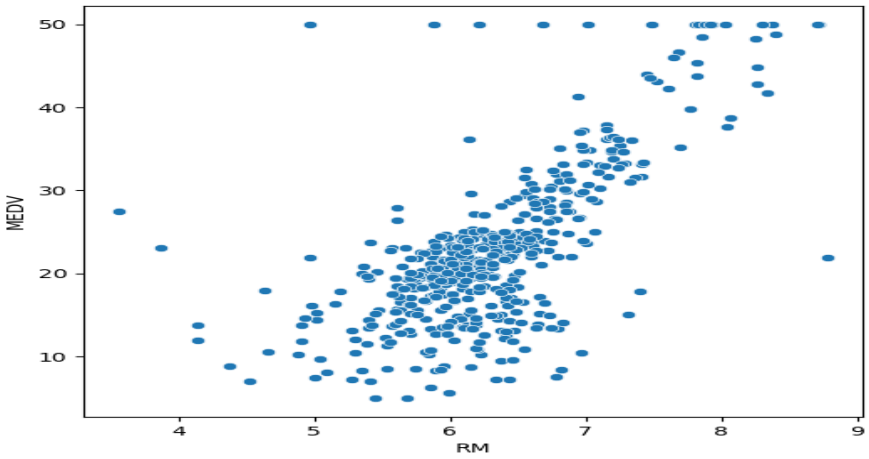
and RAD is 0.24975731331429196 after removing the outliers which means the higher property tax rate could be due to some other factors.

THE RELATIONSHIP BETWEEN INDUS AND TAX



From the above graph we see that the tax rate seems to escalate with an increase in the proportion of non-retail business acres per urban area. It could be due to possibility that the variables TAX and INDUS are related with a third variable. There exists positive relationship between INDUS (% of industrial land) and TAX (property tax rate) as shown by scatter plot. Areas with a higher percentage of industrial land normally tend to have higher property tax rates, the correlation coefficient (≈ +0.72) confirms a strong positive correlation. This recommends that industrial zoning subsidises to higher tax evaluations, perhaps as industrial regions necessitate further infrastructure, leading to higher taxation. However, the trend is evident but some variability exists as not all industrialized areas necessarily surface the same tax tariffs.

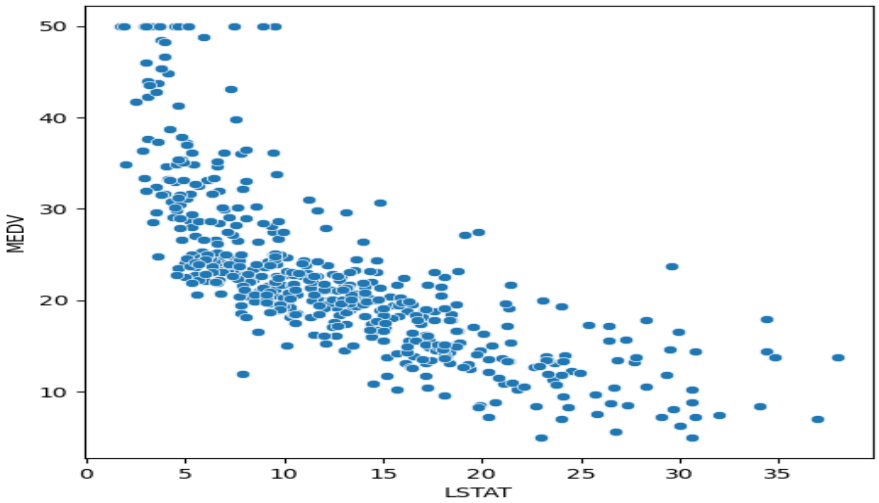
THE RELATIONSHIP BETWEEN RM AND MEDV



The scatterplot shows the value of the house which appears to intensify as the value of RM increases. This is expected as the price is generally higher for more rooms. There are a few outliers in a horizontal line as the MEDV value seems to be capped at 50. Therefore, RM and MEDV shows a strong positive corelation. As the average number of rooms per dwelling (RM) increases, the median house price (MEDV) also increases. The correlation coefficient (~ +0.7) confirms a strong linear association. Maximum points shadow an ascending trend, representing that bigger houses with more rooms are normally extra luxurious and have higher prices. The general

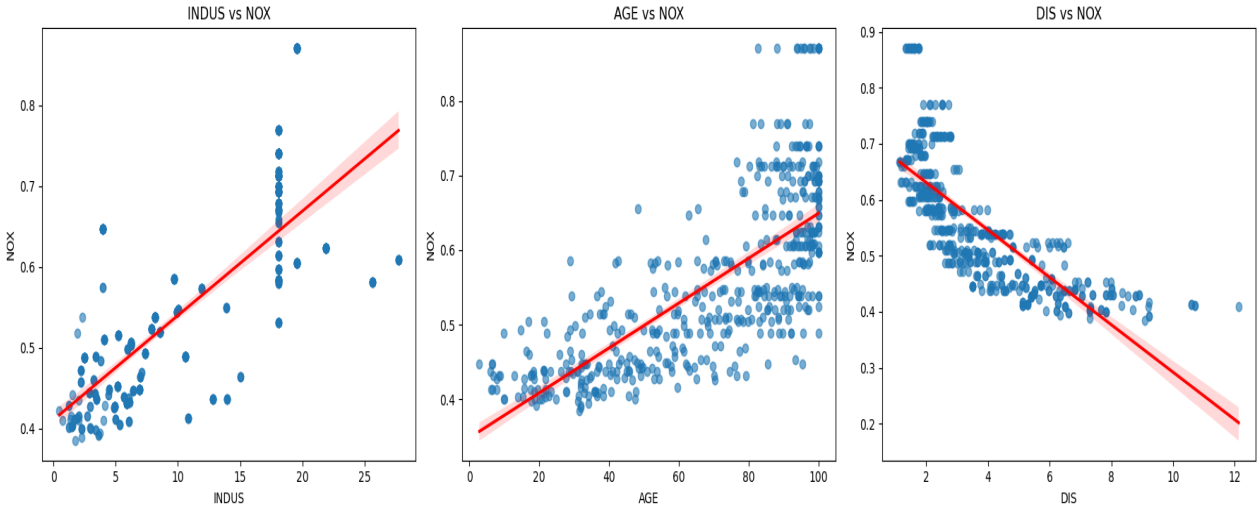
positive arrangement remains perfect as few outliers exist at very high RM values (>8). The variable RM is estimated to be one of the strongest predictors of housing prices in the regression model.

THE RELATIONSHIP BETWEEN LSTAT AND MEDV



The above scatterplot shows that as LSTAT increases, the price of the house tends to decrease which shows a strong negative relationship between LSTAT (% of lower status population) and MEDV (median home value). This might be due to the cause of lower houses prices somewhere the individual with inferior status live. There are scarce outliers and the data appears to be capped at 50. As the proportion of lower-status population (LSTAT) increases, the median house price (MEDV) decreases, also the correlation coefficient (~ –0.74) confirms a strong inverse relationship. The descending trend is reasonably perfect and reliable, yet with some scatter at higher LSTAT values. This recommends that socio-economic status strongly influences housing prices as the richer neighbourhoods (low LSTAT) have appreciably higher house prices. Beside RM, LSTAT is expected to be one of the most important predictors of MEDV in the regression model.

THE RELATIONSHIP BETWEEN INDUS VS NOX, AGE AND AGE AND NOX, DIS VS NOX:



The above three scatterplots show the relationship of INDUS, AGE, and DIS vs NOX. The correlation values for the features with strong correlations to NOX are:

1. INDUS vs NOX:

Since the fraction of non-retail business acres (INDUS) rises, the nitric oxide concentration (NOX) also rises as the correlation value is +0.764, which shows that the industrialized areas lean to have higher pollution levels. The graph shows the strong positive linear trend.

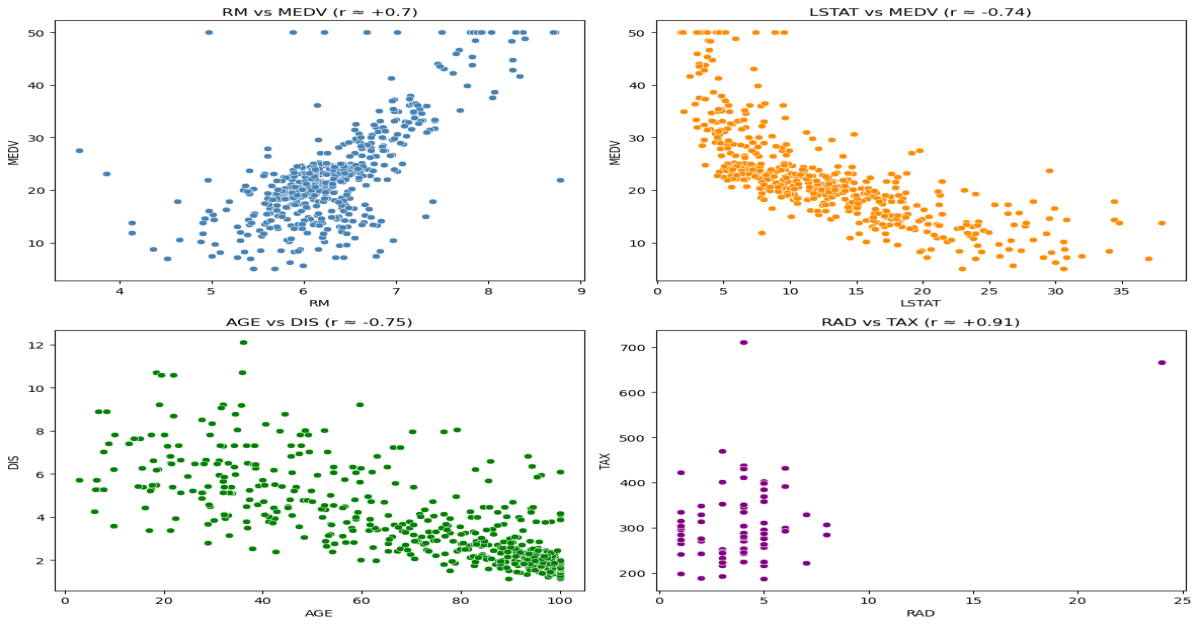
2. AGE vs NOX:

The above plot shows areas with older houses (AGE) generally demonstrate higher NOX levels with the corelation value about +0.731. This value reflects strong positive relationship and recommends that older neighbourhoods are closer to polluted/industrial regions.

3. DIS vs NOX

Following the relationship between DIS and NOX the correlation value is -0.769, it shows that the weighted distance to employment centres (DIS) has a robust negative relationship with NOX. The farther a neighbourhood is from business/industrial hubs, the lower its pollution level, hence strong negative linear trend is seen.

Concluding above, these three variables (INDUS, AGE, DIS) indicates the strong correlation with pollution levels (NOX), therefore formulating them significant predictors in regression models. Visualizing the relationship between the features having significant correlations (>= 0.7 or <= -0.7) as below:



SUMMARY:

The above analysis shows that variables with higher absolute correlation (|r| close to 1) tend to have more linear relationships with MEDV. There exists strong linear relationship of MEDV with MV and LSTAT. Strong positive upward linear relationship appears between MEDV and RM as there are more rooms, higher are the house prices. Similarly, strong negative linear trend exists between MEDV and LSTAT as percentage of lower-status population is higher, the house price would be lower. Other variables with negative impact are PTRATIO, INDUS, TAX, NOX, and the weak positive influence variables are ZN, DIS, CHAS. Hence, higher student-teacher ratio yields lower prices. More industrial area moves to lower prices. Higher property tax yields lower prices. As the higher pollution gives rise to lower prices. Moreover, if there are more residential land zone, there would be slightly higher prices. Since the distance from employment centres affect slightly higher prices (near Charles River although counted as categorical yields somewhat higher prices).

REGRESSION ANALYSIS

Once the exploratory data analysis is done, we proceed further to build a regression model of MEDV (house prices) using the important predictors. First of all, we will select features which is the strongest predictors, LSTAT, RM, also other moderately correlated features including PTRATIO, INDUS, TAX, NOX, CRIM. Secondly, we will train-test split data in which we will divide the dataset into training and testing sets in a ratio of 70:30 to evaluate performance. Next, we will separate the dependent variables and independent variables and will check the multicollinearity in that training dataset by Variance Inflation Factor (VIF), the variables having VIF score greater than 5 will be dropped and will be treated till all features have a VIF score less than 5. Further, we will fit Linear Regression to start with Simple Linear Regression (MEDV vs LSTAT, MEDV vs RM) and then move to Multiple Linear Regression with several predictors. Finally, we will evaluate model by checking R^2 score, Adjusted R^2, and RMSE to interpret coefficients and how each variable impacts the target variable MEDV.

Check for Multicollinearity

As we will use the Variance Inflation Factor (VIF), to check if there is multicollinearity in the data. Features having a VIF score greater than 5 will be dropped and will be treated till all the features have a VIF score less than 5. Simple Interpretation:

VIF < 5 = (low multicollinearity).

VIF 5–10 = Moderate multicollinearity.

VIF > 10 = High multicollinearity (in this case we should drop or transform variables).

Index	Features	VIF
0	const	535.372593
1	CRIM	1.924114
2	ZN	2.743574
3	INDUS	3.999538
4	CHAS	1.076564
5	NOX	4.396157
6	RM	1.86095
7	AGE	3.15017
8	DIS	4.355469
9	RAD	8.345247
10	TAX	10.191941
11	PTRATIO	1.943409
12	LSTAT	2.861881

Since the above table shows that there are two variables with a high VIF i.e., RAD and TAX (values greater than 5), so we will remove TAX as it has the highest VIF values and in order to check the multicollinearity again.

Index	Feature	VIF
0	const	532.025529
1	CRIM	1.923159
2	ZN	2.483399
3	INDUS	3.270983
4	CHAS	1.050708
5	NOX	4.361847
6	RM	1.857918
7	AGE	3.149005
8	DIS	4.333734
9	RAD	2.942862
10	PTRATIO	1.90975
11	LSTAT	2.860251

Conclusively, we will generate the linear regression model and get the model summary as the VIF is less than 5 for all the independent variables in the above table, and we can suppose that multicollinearity has been removed among the variables.

OLS REGRESSION MODEL AND SUMMARY

OLS Regression Results			
Dep. Variable:	MEDV_log	**R-squared:**	0.769
Model:	OLS	**Adj. R-squared:**	0.761
Method:	Least Squares	**F-statistic:**	103.3
Date:	Sat, 04 Oct 2025	**Prob (F-statistic):**	1.40e-101
Time:	03:42:45	**Log-Likelihood:**	76.596
No. Observations:	354	**AIC:**	-129.2
Df Residuals:	342	**BIC:**	-82.76
Df Model:	11		
Covariance Type:	Nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6324	0.243	19.057	0.000	4.154	5.111
CRIM	-0.0128	0.002	-7.445	0.000	-0.016	-0.009
ZN	0.0010	0.001	1.425	0.155	-0.000	0.002
INDUS	-0.0004	0.003	-0.148	0.883	-0.006	0.005
CHAS	0.1196	0.039	3.082	0.002	0.043	0.196
NOX	-1.0598	0.187	-5.675	0.000	-1.427	-0.692
RM	0.0532	0.021	2.560	0.011	0.012	0.094
AGE	0.0003	0.001	0.461	0.645	-0.001	0.002
DIS	-0.0503	0.010	-4.894	0.000	-0.071	-0.030
RAD	0.0076	0.002	3.699	0.000	0.004	0.012
PTRATIO	-0.0452	0.007	-6.659	0.000	-0.059	-0.032
LSTAT	-0.0298	0.002	-12.134	0.000	-0.035	-0.025

Omnibus:	30.699	**Durbin-Watson:**	1.923
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	83.718
Skew:	0.372	**Prob(JB):**	6.62e-19
Kurtosis:	5.263	**Cond. No.**	2.09e+03

This above model results summary provides a strong and statistically significant explanation of Boston Housing Prices (after log transformation), with good explanatory power (Adj. $R^2 = 0.761$). The above table shows OLS Regression Results summary for the Boston Housing model (MEDV_log as dependent variable). The model is best Fit, R-squared is equal to 0.769 which is about 76.9% of the variation in housing prices (log-transformed) which is described by the independent variables in the model. The value of Adjusted R-squared is equal to 0.761 after correcting for the number of forecasters, the model nevertheless explains 76.1% of the discrepancy. This illustrates the model has strong explanatory control.

Additionally, the statistical significance of the model shows the F-statistic = 103.3, Prob (F-statistic) equals to 1.40e-101. The actual little p-value confirms that the generally model is statistically significant, implicating at least one predictor variable has a significant correlation with house rates and their prices. The Log-Likelihood equals to 76.596 which indicates the model's fit. A higher value normally means a better fit, but it's primarily valuable when associating with unconventional models. The AIC value is equal to -129.2 and BIC equals to -82.76. Both are perfect range measures where lower values suggest a improves the model. These values can be compared with alternative models (for example, with fewer predictors) to check if complication is acceptable.

The observations and degrees of Freedom have 354 observations which are used in the training dataset. The df Residuals value is equal to 342 and Df model value is equal to 11. The model has estimated 11 independent predictors, leaving 342 degrees of freedom for residuals. The covariance type is non robust and the standard errors are based on classical OLS assumptions. If heteroscedasticity exists, robust errors might be needed for more reliable inference.

SUMMARY TABLE OF OLS REGRESSION

Predictor	Coefficient	p-value	Significance	Interpretation
CRIM (Crime rate)	-0.0128	0	Significant	Higher crime rates significantly reduce housing values.
CHAS (Charles River dummy)	0.1196	0.002	Significant	Houses near the Charles River are more expensive.
NOX (Nitric Oxide pollution)	-1.0598	0	Significant	Higher air pollution levels decrease prices sharply.
RM (Average rooms per dwelling)	0.0532	0.011	Significant	More rooms lead to higher housing prices.
DIS (Distance to employment centers)	-0.0503	0	Significant	Houses farther from job centres are less valuable.
RAD (Highway accessibility index)	0.0076	0	Significant	Better access to highways raises property value.
PTRATIO (Pupil-teacher ratio)	-0.0452	0	Significant	Poorer schools (higher ratio) lower property value.
LSTAT (% lower status population)	-0.0298	0	Significant	Higher proportion of low-status population reduces housing prices.
ZN (Residential land proportion)	0.001	0.155	Not significant	No strong effect once other variables are controlled.
INDUS (Non-retail business land)	-0.0004	0.883	Not significant	Industrial land proportion doesn't affect prices here.
AGE (Older houses proportion)	0.0003	0.645	Not significant	Age of housing is not statistically significant.

INTERPRETATION OF REGRESSION RESULTS

The regression analysis shows that several predictors have a statistically significant impact on housing prices. LSTAT (–0.0298) and CRIM (–0.0128) are strong negative predictors, indicating that higher crime rates and larger proportions of lower-status populations substantially reduce property values. RM (+0.0532) has a strong positive effect, meaning that homes with more rooms are significantly more expensive. Environmental quality also plays a crucial role, with NOX (–1.0598) showing that higher pollution levels sharply decrease housing prices. Educational quality, measured by PTRATIO (–0.0452), is another significant negative factor. Proximity to the Charles River (CHAS, +0.1196) adds a positive premium. Meanwhile, variables like ZN, INDUS, and AGE were not statistically significant, suggesting limited independent influence once stronger predictors are controlled for. Overall, the model highlights that socioeconomic conditions, structural features, and environmental quality are the primary drivers of Boston housing values.

CONCLUSION:

The Standard Errors assume that the covariance matrix of the errors is correctly specified. The Boston Housing project analysis shows that the house prices increase with more rooms, proximity to Charles River, better accessibility, and lower lawbreaking and less pollution smog. Whereas, the houses prices decrease with higher crime, higher pollution, greater distance to employment centres, low schools grading, and higher percentage of low status population. The condition number is large, 2.09e+03. The value reflects that there could be robust multicollinearity or some numerical values problem in the data.

ELIMINATING THE INSIGNIFICANT FEATURES AND CREATING THE NEW MODEL:

Since it is not sufficient to fit a multiple regression model to the data; it is essential to check whether all the regression coefficients are significant or not. Significance here means whether the population regression parameters are significantly different from zero. From the above analysis it may be distinguished that the regression coefficients similar to ZN, AGE, and INDUS are not statistically meaningful at level $\alpha = 0.05$. In other words, the regression coefficients corresponding to these three are not significantly dissimilar from 0 in the population. Therefore, we will exclude the three features and construct a new model.

OLS Regression Results

Dep. Variable:	MEDV_log	**R-squared:**	0.767
Model:	OLS	**Adj. R-squared:**	0.762
Method:	Least Squares	**F-statistic:**	142.1
Date:	Sat, 04 Oct 2025	**Prob (F-statistic):**	2.61e-104
Time:	04:05:31	**Log-Likelihood:**	75.486
No. Observations:	354	**AIC:**	-133.0
Df Residuals:	345	**BIC:**	-98.15
Df Model:	8		
Covariance Type:	Non robust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6494	0.242	19.242	0.000	4.174	5.125
CRIM	-0.0125	0.002	-7.349	0.000	-0.016	-0.009
CHAS	0.1198	0.039	3.093	0.002	0.044	0.196
NOX	-1.0562	0.168	-6.296	0.000	-1.386	-0.726
RM	0.0589	0.020	2.928	0.004	0.019	0.098
DIS	-0.0441	0.008	-5.561	0.000	-0.060	-0.028
RAD	0.0078	0.002	3.890	0.000	0.004	0.012
PTRATIO	-0.0485	0.006	-7.832	0.000	-0.061	-0.036
LSTAT	-0.0293	0.002	-12.949	0.000	-0.034	-0.025

Omnibus:	32.514	**Durbin-Watson:**	1.925

Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	87.354
Skew:	0.408	**Prob(JB):**	1.07e-19
Kurtosis:	5.293	**Cond. No.**	690.

SUMMARY OF THE RESULTS:

The above table shows the refined model for the Boston Housing dataset (dependent variable: MEDV_log). The model Fit and performance shows R^2 is equal to 0.767 whereas Adjusted R^2 is equal to 0.762. The model explains about 76% of the variation in log house prices. This is a very strong fit for real-world cross-sectional housing data. The F-statistic (very small p-value) shows that the model as a whole is greatly significant. The Durbin-Watson value is equal to 1.925 which recommends no serious autocorrelation in residuals (as values close to 2 are good). For all the significant predictors, $p < 0.05$, every included predictor appears to be significant. Higher crime rates reduce housing values CRIM (-0.0125), houses near the Charles River have higher prices CHAS (+0.1198), Pollution is strongly negative for housing prices NOX (-1.0562), More rooms lead to higher housing values RM (+0.0589), Greater distance from employment centres lowers housing values DIS (-0.0441), Easy access to highways increases property values RAD (+0.0078), Higher student–teacher ratios (weaker schools) reduce prices PTRATIO (-0.0485), A higher proportion of lower-status population depresses housing values strongly LSTAT (-0.0293).

For the residual diagnostics the Omnibus & Jarque-Bera tests ($p < 0.000$) where residuals are not perfectly normal; some skewness (0.408) and kurtosis (5.293), this is typical in real estate data. Condition Number equals to 690 which leads to lesser than earlier models (2000+), the results shows that now the significance multicollinearity is less severe. The key observations are that this reduced model is stronger and more parsimonious: all predictors are significant, and AIC/BIC values are enhanced. Main drivers of housing prices are crime, air quality, proximity to Charles River, number of rooms, convenience, school quality, and socio-economic factors. In the previous model, statistically insignificant variables (ZN, INDUS, AGE) were dropped, improving efficiency without dropping explanatory power.

EXECUTIVE SUMMARY OF REDUCED REGRESSION MODEL:

The reduced OLS regression model describes approximately 76% of the variation in Boston housing prices (log-transformed), demonstrating robust extrapolative power. Contrasting the full model, all eight predictors engaged are statistically significant, creating this form both parsimonious and robust. Results show that housing values increase with more rooms, proximity to the Charles River, and better highway access, while they decrease with higher crime rates, air pollution, greater distance from employment centres, low academic school quality and grading, and higher proportions of lower-status population. Compared to the earlier model that included non-significant predictors (ZN, INDUS, AGE), this model rationalised specification improves model efficiency (lower AIC/BIC) while retaining explanatory strength, providing clearer insights into the key socio-economic and environmental drivers of housing prices.

LINEAR REGRESSION ASSUMPTIONS

1. Mean of residuals should be 0
2. No Heteroscedasticity
3. Linearity of variables
4. Normality of error terms

RESIDUALS:

The mean of residuals for your OLS regression model is approximately -0.01478. Interpreting, this value is very close to 0, which satisfies the regression assumption that residuals should have a mean of 0. The small negative bias (−0.0147) indicates the model's predictions are, on average, only slightly overestimating house prices (in log scale).

HOMOSECDASTICITY GOLDFELD-QUANDT TEST AND HETROSCEDASTICITY:

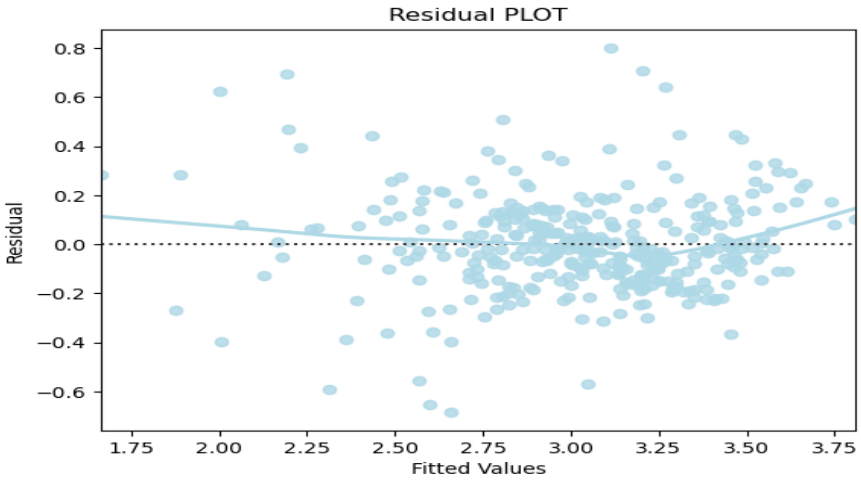
The test used is Goldfeld–Quandt test:

Null Hypothesis (H_0): Residuals are homoscedastic (constant variance)
Alternate Hypothesis (H_1): Residuals are heteroscedastic (non-constant variance)
F-statistic value = 1.0835 and the P-value = 0.3019

The observation shows that since the p-value > 0.05, we fail to reject the null hypothesis. This means the residuals are homoscedastic, i.e., the variance of the error terms is constant across observations. Therefore, the homoscedasticity assumption of linear regression is satisfied for the model. (If actual p-value ≤ 0.05, then the interpretation flips: we reject H_0 and conclude that the data suffers from heteroscedasticity.

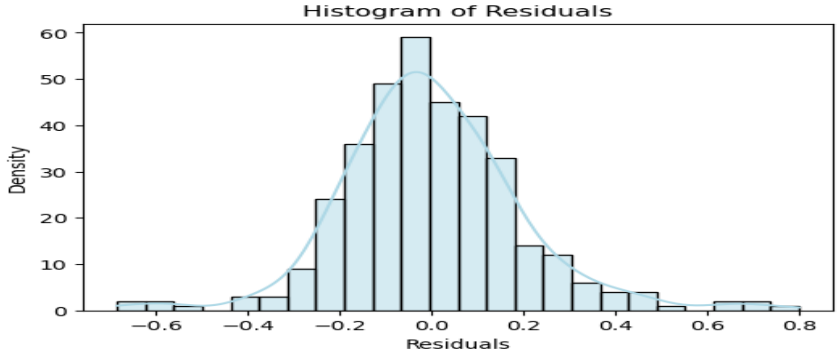
LINEARITY OF VARIABLES:

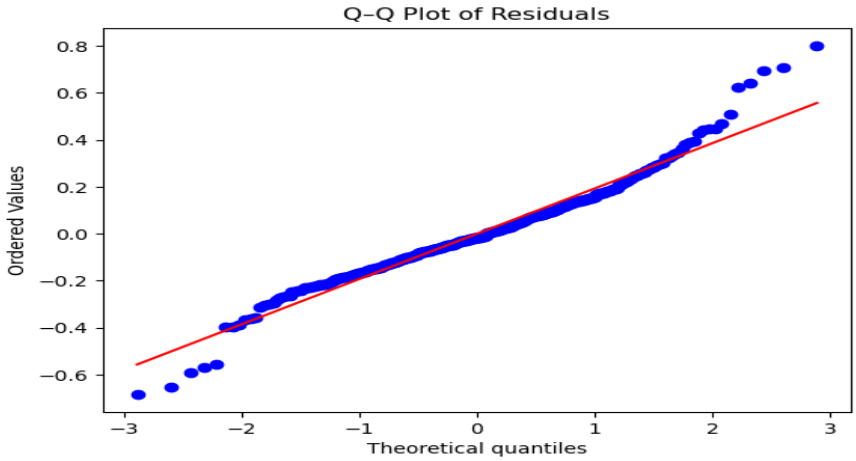
It states that the predictor variables must have a linear relation with the dependent variable. To test the assumption, we'll plot residuals and the fitted values on a plot and ensure that residuals do not form a strong pattern. They should be randomly and uniformly scattered on the x-axis.



The linearity check (Residual vs Fitted plot) gives the observation that in the Mean of Residuals, the residuals are centred around zero, which confirms that the average residual is approximately 0 which satisfies one of the key assumptions of linear regression. Further, for the linearity of Variables, from the Residual vs Fitted plot, if the points are randomly scattered around the horizontal line (y = 0) without a clear pattern, it suggests that the relationship between predictors and MEDV log is linear. If a curve or systematic shape is visible, that would mean some non-linearity is present, and we may need polynomial/interaction terms.

NORMALITY OF ERROR TERMS:





From the above graphs we see that the shape of the histogram of residuals is approximately bell-shaped and centred around 0, indicating residuals are roughly symmetric. The Q–Q plot shows that the points lie close to the 45° line, with only minor deviations at the tails which shows that the residuals are approximately normal. If p-value > 0.05 we fail to reject as normality assumption satisfied. If p-value ≤ 0.05 there will be some departure from normality, hence in practice OLS is fairly robust, but we can consider transforming variables, removing outliers, or using robust/bootstrapped inference if needed. Overall, the normality of error terms assumption appears reasonably satisfied for this model. The histogram of residuals is approximately bell-shaped and centred around 0, which shows that the residuals are roughly symmetric. In the Q–Q Plot, most points lie close to the 45° reference line, but there are visible deviations in the tails. This indicates that residuals follow a near-normal distribution but with some outliers/extremes. The Jarque–Bera Test Results shows JB = 87.354, p = 0.0000, skew = 0.408, kurtosis = 5.293. Since p < 0.05, we reject the null hypothesis of perfect normality. The skew is small (0.408 → mild right skew), but kurtosis is high (>3), suggesting heavy tails (leptokurtosis).

CONCLUSION:

The residuals are approximately normal but not perfectly so. The main departure comes from heavy-tailed residuals (outliers at the extremes). Despite this, linear regression is robust to mild violations of normality, so the model is still reliable for inference and prediction, but p-values and confidence intervals should be interpreted with caution.

Results from Residual vs Fitted Plot: The mean of residuals is approximately –0.015, which is very close to 0. The residuals are scattered randomly around the zero line across the fitted values, with no obvious curve or trend. The variance of residuals appears fairly constant, indicating homoscedasticity. A few points lie farther away from zero, suggesting possible outliers, but they do not show a systematic pattern.

Results from Normality Checks (Histogram & Q-Q Plot): Histogram: Residuals follow an approximately bell-shaped distribution, close to normal. Q-Q Plot: Most points lie along the 45° reference line, with only minor deviations at the tails. This confirms that the normality of error terms assumption holds reasonably well.

Finishing, Linearity is maintained since residuals vs fitted plot shows random scatter. Homoscedasticity is sustained as no funnel-shaped pattern, variance is stable. Normality of residuals is also maintained since histogram and Q-Q plot confirm approximate normality. Lastly, the mean of residuals is close to zero (–0.015), which is expected. Hence, the regression assumptions are fulfilled, so model is statistically valid for inference and prediction.

PERFORMANCE OF THE MODEL ON TEST AND TRAIN DATA:

Now, checking the performance of the model on the train and test data set

Data	RMSE	MAE	MAPE
0 Train	0.195504	0.143686	4.981813
1 Test	0.198045	0.151284	5.257965

Interpreting the results the Train and Test errors are nearly identical therefore no overfitting, the model generalizes well. Accuracy (log scale) RMSE similar to 0.20 and MAE similar to 0.15 on the log (MEDV) scale indicate small average deviations in log prices. A MAPE nearly 5% means, on average, predictions are within 5% of the actual (log) value, very good for this dataset. The key notes are that the model is stable and accurate after multicollinearity cleanup (dropping TAX, ZN, AGE, INDUS). Errors on the unobserved data are only marginally higher than on training data, confirming robustness. So, errors are almost identical between training and test data so this shows no overfitting and endorses that the model generalizes well to unseen data. The linear regression model (after handling multicollinearity and log-transforming MEDV) performs strongly, with moderate error levels in the original scale and stable generalization across train and test data.

APPLYING CROSS VALIDATION TO IMPROVE THE MODEL:

Now applying cross-validation on Boston dataset with log (MEDV) and evaluating it with multiple metrics. This way, we will get a full view of model generalization. The value of R Squared is 0.729 (+/- 0.232) and the Mean Squared Error is 0.041 (+/- 0.023). The model explains about 73% of the variance in log (MEDV), which is quite good for real-world housing data. Standard deviation shows some variability across folds, but the model is reasonably consistent. MSE (0.041 in log scale), in log terms, this error is small. Back-transforming: RMSE is approximately equal to $\sqrt{0.041}$ equals to 0.202 as in original house prices, this is like about 20% error on average (approximately equal to $3K–$4K if median home is about $22K in dataset scale). MAE (0.155 in log scale) this means the average absolute error in log (MEDV) is about 0.155. When back-transformed, this corresponds to about 15% typical error in predicting home prices.

CONCLUSION:

Cross-validation shows that your linear regression model is robust and generalizes well after using log transformation. The average error corresponds to a few thousand dollars, which is quite acceptable for housing price prediction.

MODEL COEFFICIENTS SUMMARY:

	Feature	Coefs
0	const	4.649386
1	CRIM	-0.012500
2	CHAS	0.119773
3	NOX	-1.056225
4	RM	0.058907
5	DIS	-0.044069
6	RAD	0.007848
7	PTRATIO	-0.048504
8	LSTAT	-0.029277

Here are the model coefficients from regression (with MEDV_log as dependent variable). The feature Coefficient values are as const 4.6494 CRIM -0.0125 CHAS +0.1198 NOX -1.0562 RM +0.0589 DIS -0.0441 RAD +0.0078 PTRATIO -0.0485 LSTAT -0.0293. Interpretating the coefficients intercept (const = 4.65), the baseline log house price when all features are equal to zero. For CRIM (-0.0125), higher crime rate is related with lower house prices. For the CHAS (+0.1198), houses near the Charles River tend to have greater prices (positive premium effect). For the variable NOX (-1.0562), higher air pollution (NOX concentration) strongly reduces house prices. In case of RM (+0.0589), each additional average room per dwelling increases log house prices, i.e. larger houses are more expensive. The variable DIS (-0.0441), greater distance from employment centres decreases house prices. Also, in the case of RAD (+0.0078), availability to circular freeways demonstrates a small positive outcome. Variable PTRATIO (-0.0485), higher student–teacher ratios (worse school quality) reduce house prices. Lastly, for LSTAT (-0.0293) as the higher percentage of lower-status population significantly lowers house prices.

CONCLUSION:

The regression model explains a significant portion of the variance in housing prices ($R^2 \approx 0.73–0.75$ across cross-validation). The utmost prominent dynamics on house prices, the optimistic drivers, which include the number of rooms (RM), proximity to Charles River (CHAS), accessibility to radial highways (RAD). The adverse drivers are Pollution levels (NOX), distance from employment centres (DIS), crime rate (CRIM), poor student–teacher ratios (PTRATIO), and socio-economic disadvantage (LSTAT). The model's error levels are adequate, expectations are within ~$3k–$4.5k of actual values, or ~14–15% error rate, showing strong predictive consistency. The regression assumptions of Normality and homoscedasticity are reasonably satisfied, though heavy tails in residuals suggest some outlier outcomes. The model authorizes fundamental economic perception i.e., healthier environment, bigger homes, superior schools, and river closeness drive prices up, while crime, pollution, distance, and socio-economic disadvantage drive prices down.

The regression analysis of the Boston Housing dataset provides clear insights into the main drivers of property values. The results suggest that increasing the number of rooms, improving environmental quality, lowering crime rates, and heightening education (poorer pupil–teacher ratios) have the sturdiest positive results on housing prices. Businesses companies in real estate economics and finance can use these findings to direct investment and assets approaches, property valuation, estimation, growth of real property, and lending decisions, however policymakers can emphasize on urban planning and development, pollution control, and educational progresses to support environmental development in housing markets.

By leveraging this study, stakeholders and investors can make data-driven evaluations and decisions that diminishes improbability, recognize profitable opportunities and prosperous activities, and adopt socioeconomic challenges. Ultimately, the model enhances business success by improving forecasting accuracy, strengthening competitive positioning, and associating strategies with key market determinants of housing demand and affordability.

BUSINESS RECOMMENDATIONS:

Urban Development: Invest in reducing air pollution (NOX), reducing crime rates (CRIM), as these strongly decreases housing prices. Developing better public transport system and its accessibility will shrink the undesirable effect of distance (DIS).

Better Housing Policy & Planning: Encourage construction of larger homes (higher RM), which strongly increase housing value. Promote development near natural amenities (e.g., riverside areas), as proximity to the Charles River has a premium effect.

Investments in Education: Enhance school quality by improving student–teacher ratios (PTRATIO), which directly impacts housing demand and values.

Socio-Economic Agendas: Address LSTAT effects (lower socio-economic status concentration) through community uplift programs, as these areas see lower housing demand and prices.

Market Policy: Real estate developers can target areas with low crime, clean air, and larger homes to maximize value. Buyers may see opportunities in undervalued areas where infrastructure (schools, pollution control, accessibility) is improving. Hence, the model is statistically robust and postulates actionable insights for policymakers, govt officials, urban planners, and real estate stakeholders and investors.