

# **Boston House Price Prediction**

**Foundations for Data Science**

**29.07.2024**

# Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview
- EDA and Data Preprocessing
- Linear Regression and its Assumptions
- Appendix

# Executive Summary

## Problem summary:

The primary objective of this project was to predict housing prices in various suburbs or towns in Boston based on specific locality features. This involved addressing two critical questions:

1. **Accuracy of Predictions:** How accurately can we predict housing prices based on the provided criteria?
2. **Influential Factors:** What are the key factors that significantly influence housing prices?

By answering these questions, the client will gain valuable insights into potential future investment returns and will be better equipped to identify promising opportunities as well as potential risks.

To achieve the objective, I have analyzed data that included information about crime statistics, residential lands zoned for lots, presence of industrial businesses, Charles River in the neighbor, air pollution, average rooms number in properties, buildings age, distances to Boston employment centers, accessibility to radial highways, property-tax rates, pupil-teacher ratios, percentage of lower status population in the area and median value of homes occupied by owners.

The dataset included approximately 500 records. To complete the task, I developed several regression models to find the one that fulfills the following goals:

- **Accurate Predictions:** Predict housing prices for unseen data with the highest possible accuracy.
- **Variable Influence:** Understand how different variables affect housing prices.

To evaluate the models' performance, I used metrics like RMSE, MAE, MPAE, and  $R^2$ . This allowed me to compare different solutions and select the best one. I have also implemented cross-validation techniques to check if the selected models perform well on new data.

## Conclusions:

### 1. Target High-Impact Features:

- **LSTAT (Percentage of Lower Status Population):** This feature has the highest negative impact on MEDV. Reducing the percentage of lower status residents through community programs, education, and economic development could significantly increase property values.
- **RM (Average Number of Rooms per Dwelling):** This has the highest positive impact. Encouraging or facilitating the construction of homes with more rooms could boost property values.
- **NOX (Nitric Oxides Concentration):** This has a substantial negative impact. Efforts to reduce pollution and improve air quality could positively influence property values.

- **CRIM (Per Capita Crime Rate):** This negatively impacts property values. Implementing effective crime reduction strategies could lead to higher property values.

## 2. Monitor Moderate-Impact Features:

- **DIS (Weighted Distance to Employment Centers):** While the impact is moderate and negative, improving transportation infrastructure to reduce travel time to employment hubs could have a beneficial effect.
- **PTRATIO (Pupil-Teacher Ratio):** This also has a moderate negative impact. Enhancing educational resources and reducing the pupil-teacher ratio in schools could improve property values.
- **RAD (Index of Accessibility to Radial Highways):** This feature has a moderate impact and is slightly positive. Ensuring good accessibility while balancing other factors like noise pollution could be beneficial.

## 3. Negligible Impact Features:

- **CHAS (Proximity to Charles River):** This feature has a negligible impact on property values. Investments related to the proximity to the river may not yield significant returns in terms of property value increase.

## Recommendations:

### 1. Community:

- Implement socio-economic improvement programs targeting lower-status populations to enhance their living conditions.
- Invest in local businesses, create job opportunities and indirectly increase property values.

### 2. Housing Development:

- Promote the construction of larger homes with more rooms to attract higher home values by attracting higher-income residents.
- Balance development by including affordable housing to ensure a diverse and inclusive community.

### 3. Environmental Policies:

- Work with local government to strengthen environmental regulations to reduce NOX emissions and improve air quality contributing to higher property values.
- Promote green energy initiatives.
- Increase the number of parks and green spaces to improve air quality and make neighborhoods more attractive.

### 4. Educational Investments:

- Advocate for increased funding for schools to lower the pupil-teacher ratio, thereby making the area more attractive for families and increasing property values.

- Support local schools with extracurricular programs and tutoring services to enhance educational outcomes.

**5. Crime Reduction:**

- Develop comprehensive crime reduction strategies to improve community safety and attractiveness.
- Create environments that discourage criminal activity through urban planning, invest in security systems, improve street lighting
- Implementing programs aimed at reducing crime rates.

**6. Transportation and Infrastructure:**

- Invest in transportation infrastructure to reduce commute times to employment centers and improve accessibility to highways, enhancing the desirability of the area.
- Invest in better road maintenance and expanding public transport routes.

**7. Resource Allocation:**

- Focus resources on impactful areas such as socio-economic improvement, housing development, education, and environmental quality rather than on proximity to the river, which has a negligible impact.
- On the other hand, it is important to recognize the potential benefits offered by natural surroundings. Promoting the development of recreational and commercial amenities near the Charles River can be beneficial for enhancing the attractiveness of nearby properties.

# Business Problem Overview and Solution Approach

The primary objective of this project was to predict housing prices in various suburbs or towns in Boston based on specific locality features. This involved addressing two critical questions:

1. **Accuracy of Predictions:** How accurately can we predict housing prices based on the provided criteria?
2. **Influential Factors:** What are the key factors that significantly influence housing prices?

By answering these questions, the client will gain valuable insights into potential future investment returns and will be better equipped to identify promising opportunities as well as potential risks.

To achieve the objective, I have analyzed data that included information about crime statistics, residential lands zoned for lots, presence of industrial businesses, Charles River in the neighbor, air pollution, average rooms number in properties, buildings age, distances to Boston employment centers, accessibility to radial highways, property-tax rates, pupil-teacher ratios, percentage of lower status population in the area and median value of homes occupied by owners.

The dataset included approximately 500 records. To complete the task, I developed several regression models to find the one that fulfills the following goals:

- **Accurate Predictions:** Predict housing prices for unseen data with the highest possible accuracy.
- **Variable Influence:** Understand how different variables affect housing prices.

To evaluate the models' performance, I used metrics like RMSE, MAE, MPAE, and  $R^2$ . This allowed me to compare different solutions and select the best one. I have also implemented cross-validation technique to check if the selected model performs well on new data.

The results of this analysis should provide the client with valuable insights into their business targets. These insights can directly translate into more effective targeting of future investments, such as identifying desirable investment target groups, and may lead to improved risk management strategies.

However, it is important to note that the data used for model development is from 1970. Given the significant changes in the economic environment since then, the relationships and the impact of features on predicted house prices may differ in the current context. Therefore, it is crucial to continuously monitor the performance of the developed model and implement modifications as necessary to ensure its ongoing relevance and accuracy.

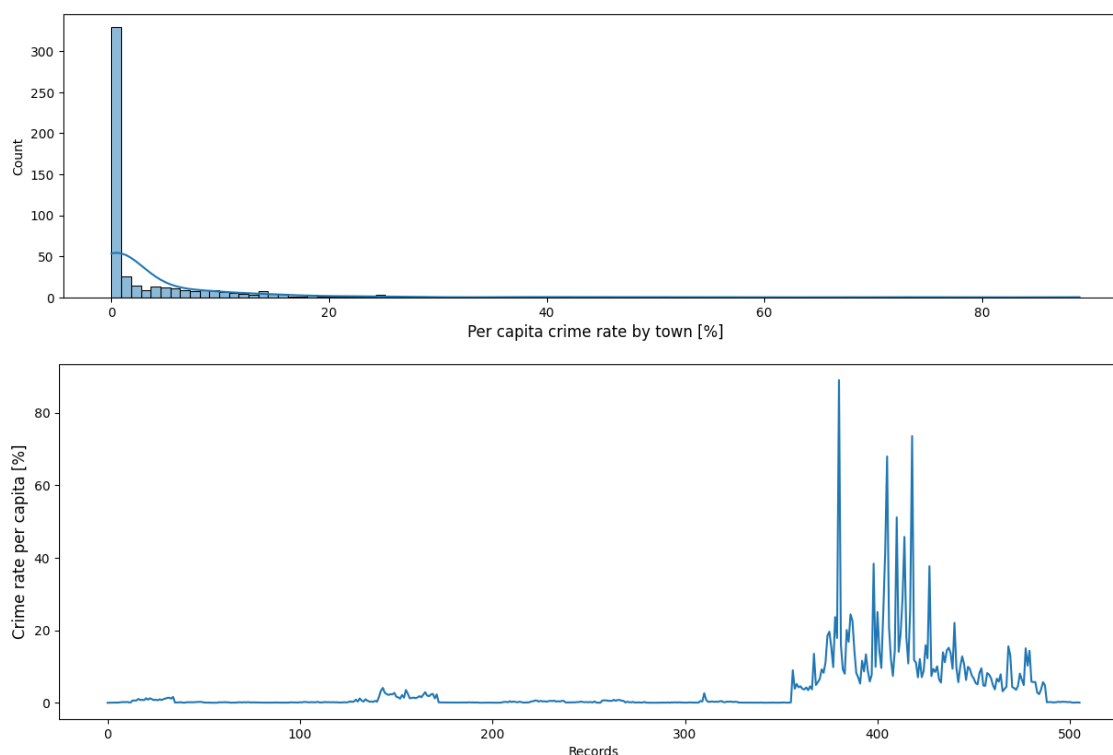
## Data Overview

- The dataset consists of 506 records with 13 columns. Each record corresponds to a suburb or town in Boston.
- Column names are without typos and white spaces.
- By calling the `info()` method we can see that there are no NaN values and data in each column is formatted correctly – numeric data type.
- There are no duplicated records in the dataset.

# EDA and Data Preprocessing

## Univariate Analysis

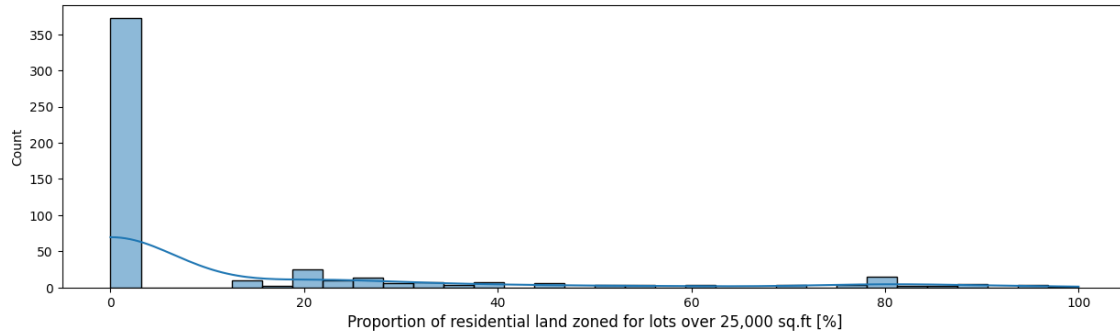
### 1. Crime rate per capita (CRIM)



- 65% of areas have CRIM less than 1%
- 75% of areas have CRIM less than 4%
- There are several areas with high CRIM values (over 10%). This factor may have a significant influence on the prices of properties. This should be examined in bivariate analysis.
- The dataset is from 1970 – Vietnam War. Maybe high crime rates are related to these actions. It therefore can be a one-off case which may alter the final predictions. We also do not know what kind of crimes comprise to the statistics
- The ten largest crime values are in the range 26% - 89% which corresponds to around 2% of the whole dataset.
- To have a better understanding of high values origin we should delve into the subject and after that decide whether to remove the outliers or leave them.



## 2. A proportion of residential land zoned for lots over 25,000 ft<sup>2</sup>. (ZN)

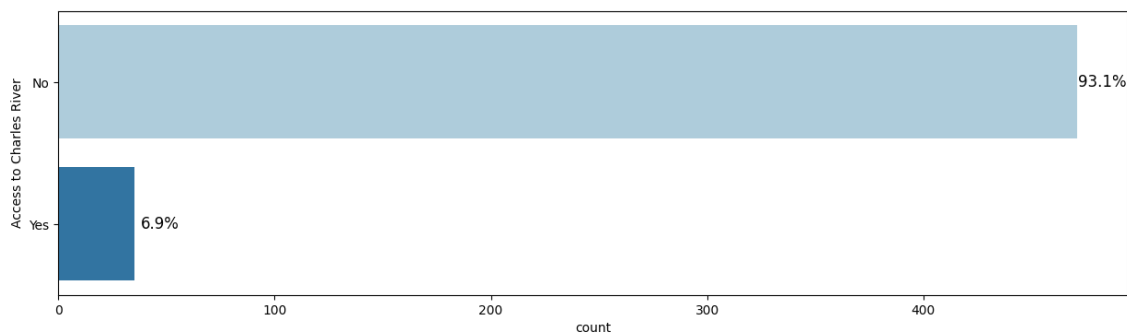


- Around 74% of areas are not intended for large residential lots.
- Only around 1% of lands is intended to have large residential lots in 90% of their area.
- Also, in some towns the whole land is dedicated to such constructions.

## 3. Proportion of non-retail business acres per town (INDUS)

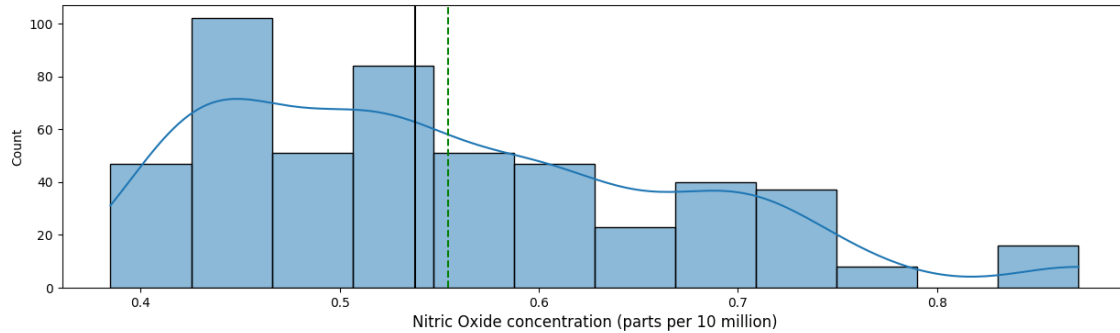
- There is a big variation of industrial areas.
- The highest proportion do not exceed 30%
- Places where nearly none of the land is dedicated to industry (less than 1%) can be found, however they are in minority (0.4%)
- Around 32% of lands have dedicated 18% - 20% of their areas for industry businesses.

## 4. Charles River (CHAS)



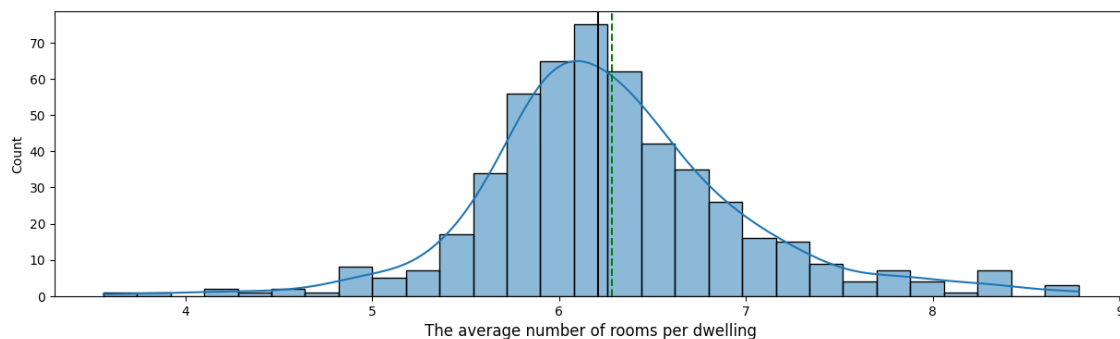
- Most of the lands do not have direct access to the Charles River.

## 5. Nitric Oxide concentration (NOX)



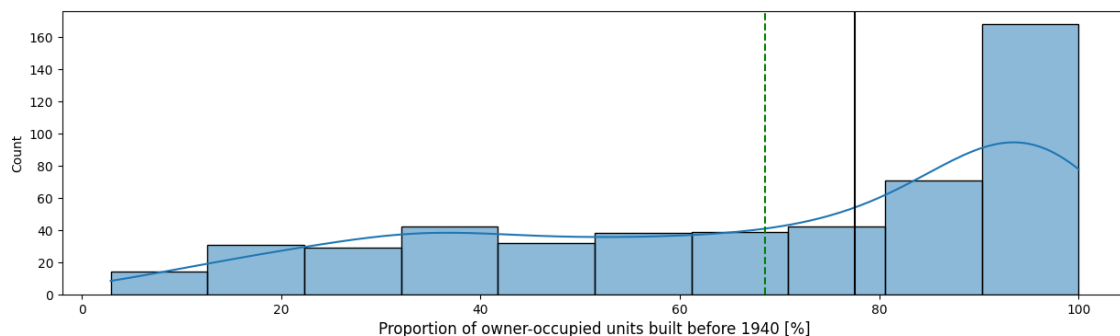
- Relatively low mean value indicates that air quality is quite good across towns.
- There are, however, areas where air pollution is high. This may be correlated with industrial zones or higher traffic.

## 6. Average number of rooms per dwelling (RM)



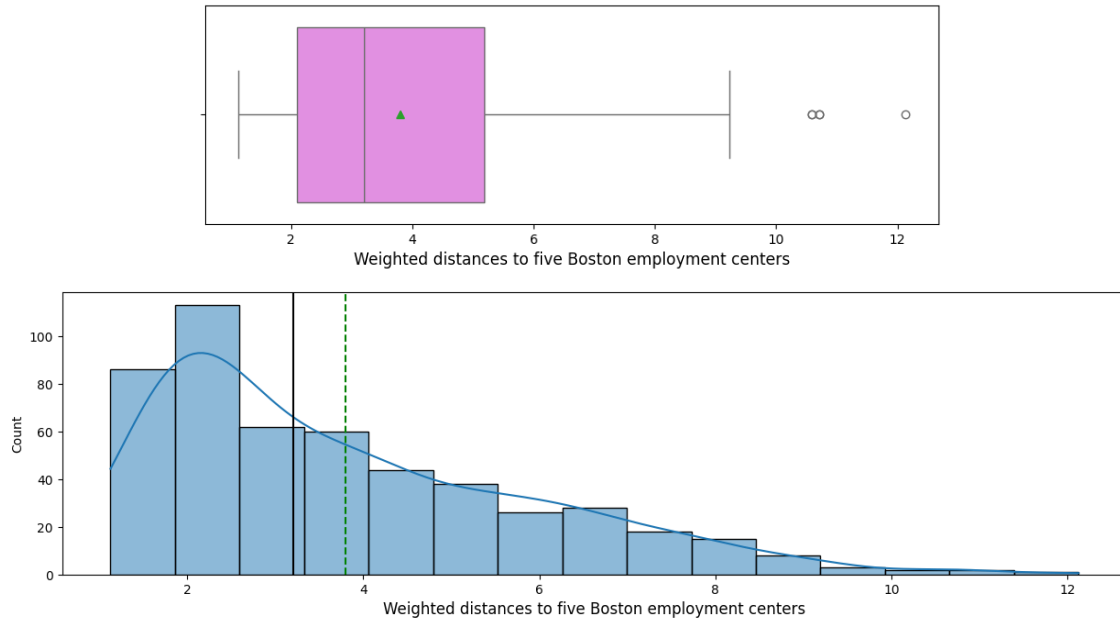
- In most towns around 95% of dwellings have 5 – 8 rooms on average.

## 7. Proportion of owner-occupied units built before 1940 (AGE)



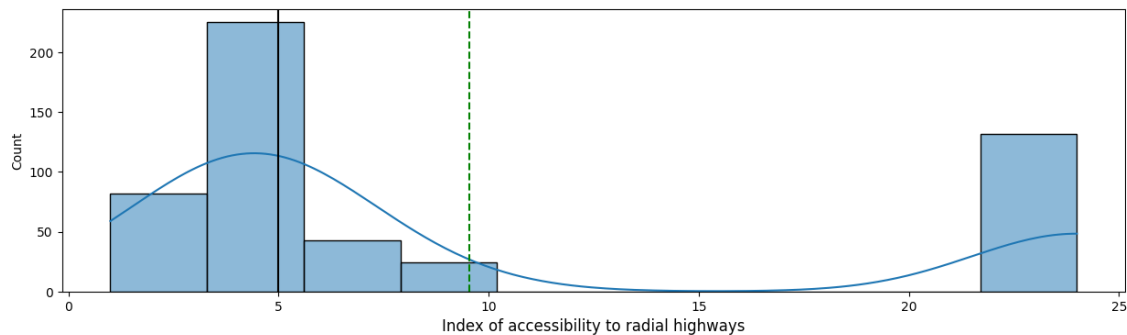
- Green dashed line represents the mean value. The black solid line stands for the median value. Both indicators levels are high which means that many homes are older than 30 years.
- In around 8.5% of lands all households were built before 1940.

## 8. Weighted distances to five Boston employment centers (DIS)

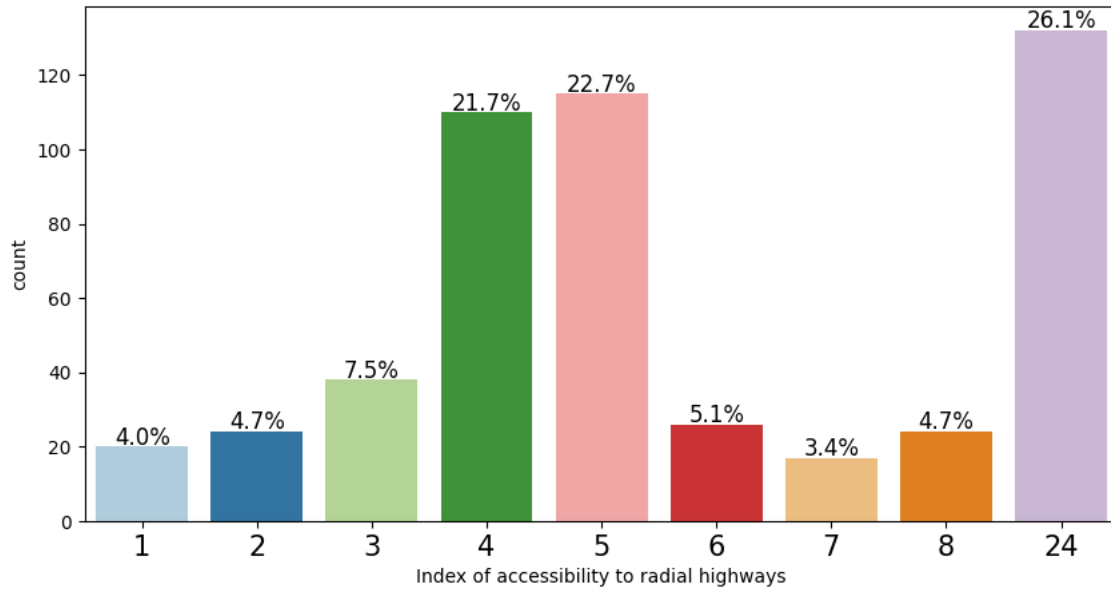


- The distribution is right skewed, which indicates that most households are in relatively close distance to the employment centers.
- There are 3 outliers, however I will not drop them as they do not seem to be wrong data.

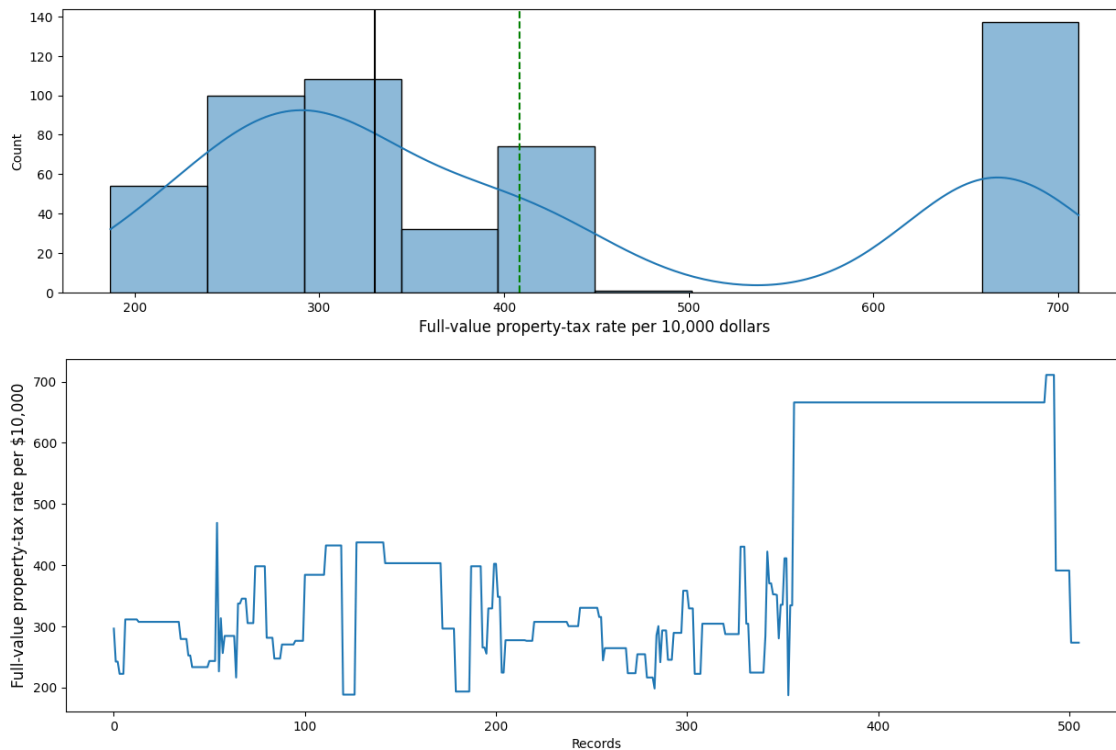
## 9. Index of accessibility to radial highways (RAD)



- There are two modes in the distribution – for RAD indices 4 and 5 (44% of lands) and for RAD index 24 (26% of lands).
- There are also areas with poor access to the highways (RAD = 1), which contribute to 4% of all data.

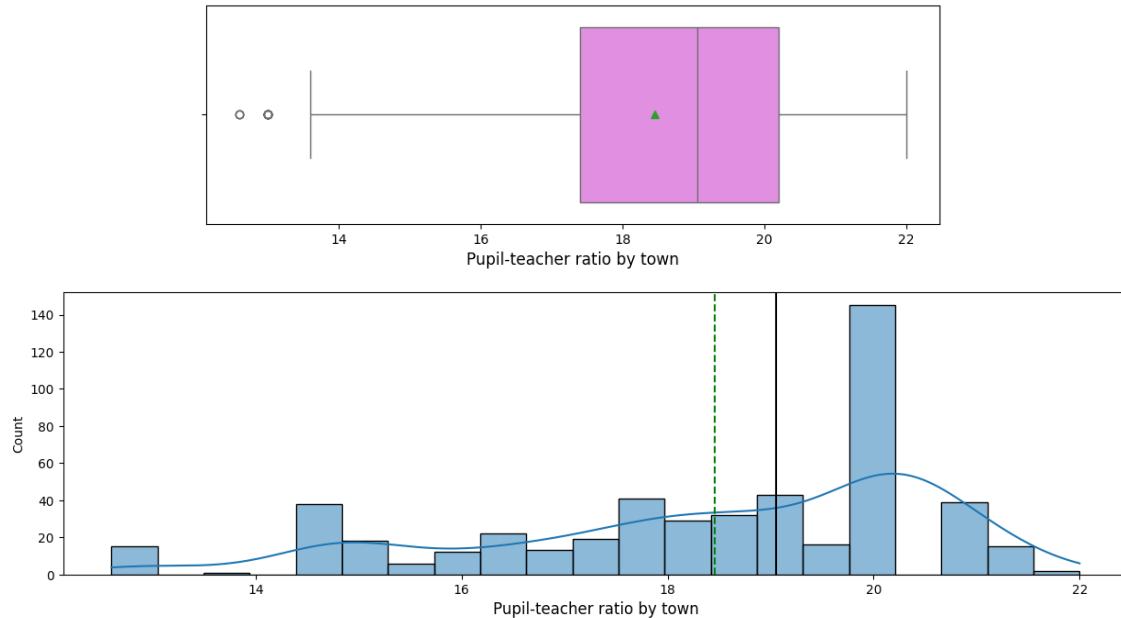


## 10. Full-value property-tax rate per 10,000 dollars (TAX)



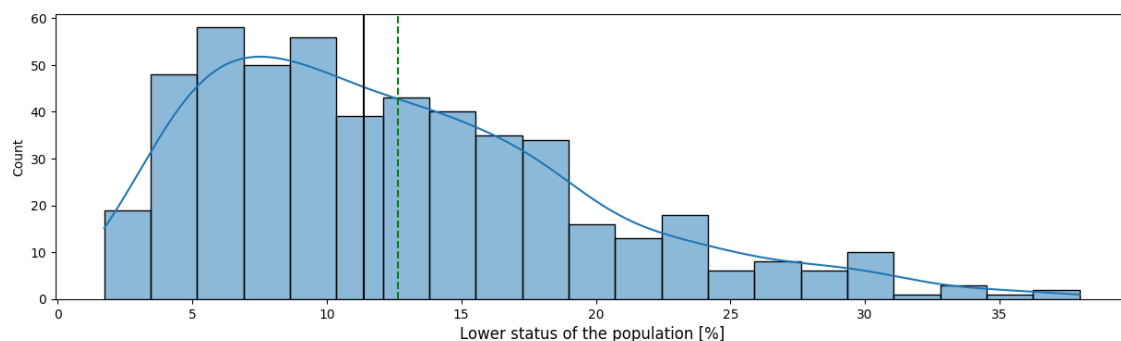
- There are around 27% of land where properties tax rates are in the range \$666 - \$711 (per \$10 000), out of which 96% have tax rates equal to \$666. This seems to be quite a lot of properties with the same tax value.
- There are also a lot of lands (44%) where properties tax rates are in the range \$300 - \$500 (per \$10 000).

## 11. Pupil-teacher ratio by town (PTRATIO)



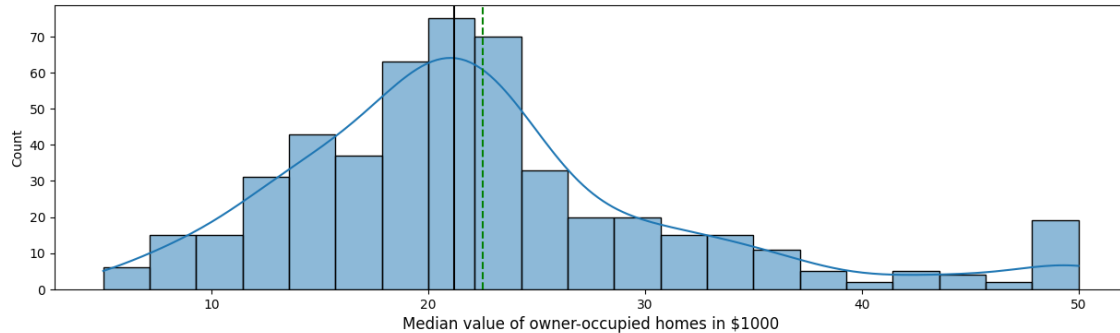
- In around half of the towns the PTRATIO is in the range of 17 – 20.
- In nearly 28% of locations there is one teacher for every 20 pupils.
- There are 15 outliers with low PTRATIO values, namely 12.6 and 13. This may indicate that there are private schools in the neighborhood.

## 12. The percentage lower status of the population (LSTAT)

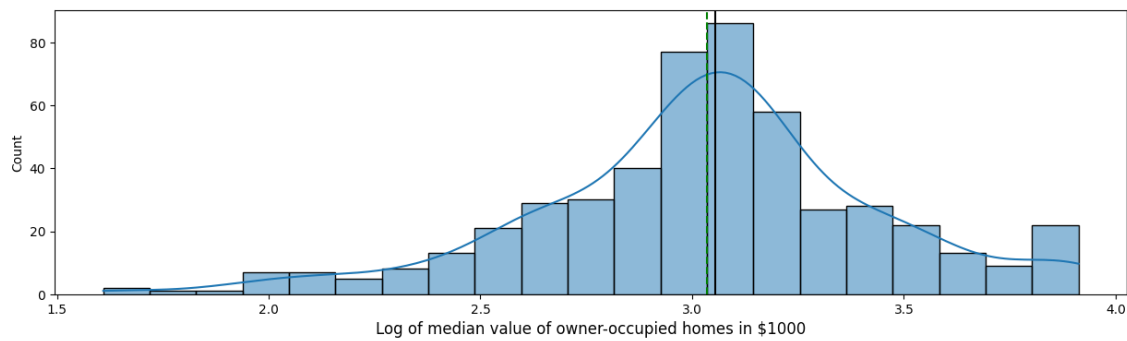


- The distribution is right skewed, in most cases lower status population does not exceed 20%.
- At some Boston areas the percentage of lower status population is high – 38%.
- This factor may have an influence on the properties price – people with lower status tend to live in cheaper houses.

### 13. Median value of owner-occupied homes in \$1000 (MEDV)



- As the dependent variable is slightly skewed, a log transformation had to be applied on the 'MEDV' values.



- The log-transformed variable appears to have a nearly normal distribution without skew.
- For the final price estimation with the developed model, we must remember that the result is on the log scale. There for the log-inverse operation (exp) must be applied to know the real predicted values.

### 14. Univariate analysis conclusions

- Univariate analysis revealed some interesting insights into the variables.
- One of the most important findings was the benefit of a log transformation applied on the dependent variable MEDV.
- The understanding gained about the behavior of independent variables will lead to a more efficient bivariate analysis.

# Bivariate Analysis

## 1. Heatmap

CRIM	1.00	-0.20	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	0.46	-0.39
ZN	-0.20	1.00	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	-0.41	0.36
INDUS	0.41	-0.53	1.00	0.06	0.76	-0.39	0.64	-0.71	0.60	0.72	0.38	0.60	-0.48
CHAS	-0.06	-0.04	0.06	1.00	0.09	0.09	0.09	-0.10	-0.01	-0.04	-0.12	-0.05	0.18
NOX	0.42	-0.52	0.76	0.09	1.00	-0.30	0.73	-0.77	0.61	0.67	0.19	0.59	-0.43
RM	-0.22	0.31	-0.39	0.09	-0.30	1.00	-0.24	0.21	-0.21	-0.29	-0.36	-0.61	0.70
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24	1.00	-0.75	0.46	0.51	0.26	0.60	-0.38
DIS	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75	1.00	-0.49	-0.53	-0.23	-0.50	0.25
RAD	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49	1.00	0.91	0.46	0.49	-0.38
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1.00	0.46	0.54	-0.47
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1.00	0.37	-0.51
LSTAT	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	1.00	-0.74
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	-0.74	1.00
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV

Positive correlations:

- The highest positive correlation is between **TAX & RAD: 0.91**
- High positive correlation is observed between **NOX & INDUS: 0.76**, which is expected.
- Other highly positive correlated values are:
  - AGE & NOX: 0.73,
  - TAX & INDUS: 0.72,
  - RM & MEDV: 0.7,

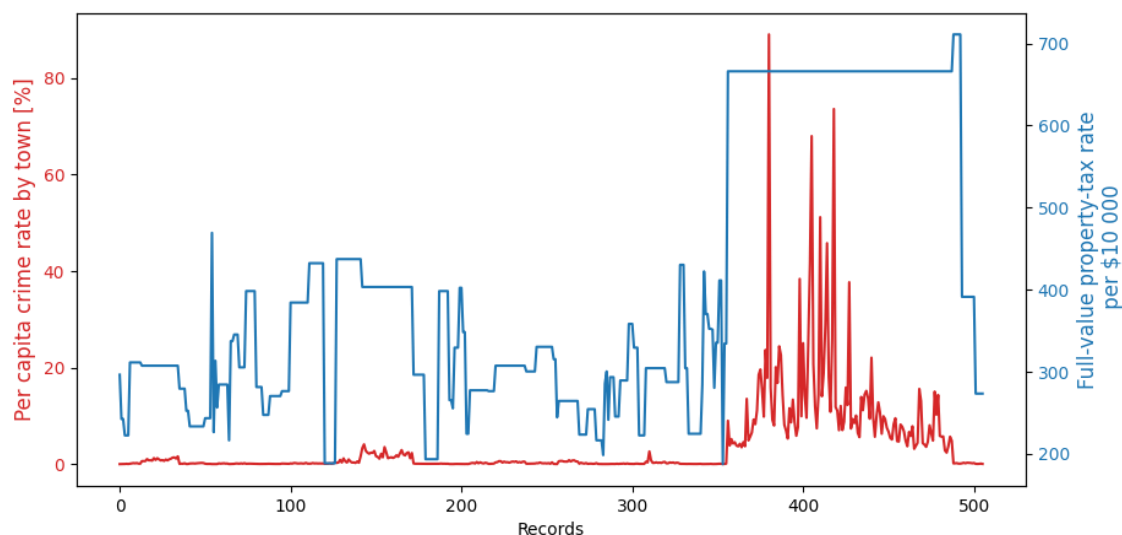
Negative correlations:

- The highest negative correlation is between **NOX & DIS: -0.77**. This may suggest that 5 Boston employment centers are from the industrial sector.
- High negative correlations can also be observed between:
  - AGE & DIS: -0.75,
  - LSTAT & MEDV: -0.74,
  - INDUS & DIS: -0.71

Other correlations:

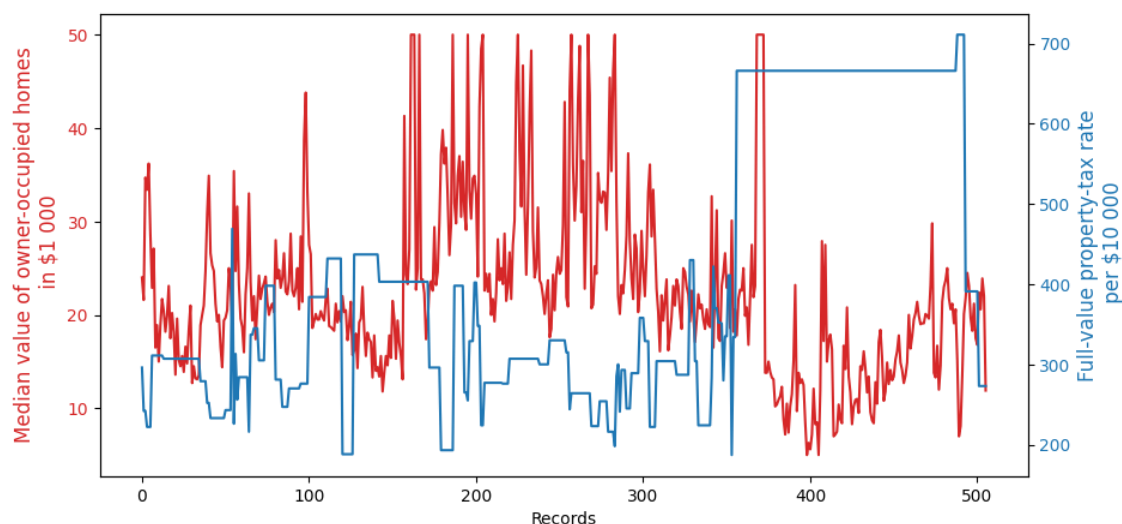
- The presence of the river basically has no correlation with any of other features (very low correlation with MEDV: 0.18)

## 2. CRIM vs. TAX



- Even though this pair has moderate positive correlation (0.58), the line plot shows that lands with high property taxes suffer from high crime rate as well.

## 3. MEDV vs. TAX



- This pair has moderate negative correlation (-0.41).
- The interesting thing is that the lands with the highest property tax rates do not correspond to those with the highest median house prices. This suggests that factors other than house

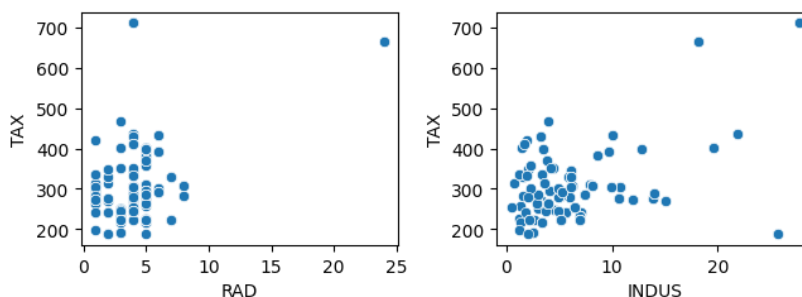


prices also influence full-value property rates, e.g. land values (high values for lands located in the city center).

#### 4. Relationships between features of correlation factor $\geq |0.7|$

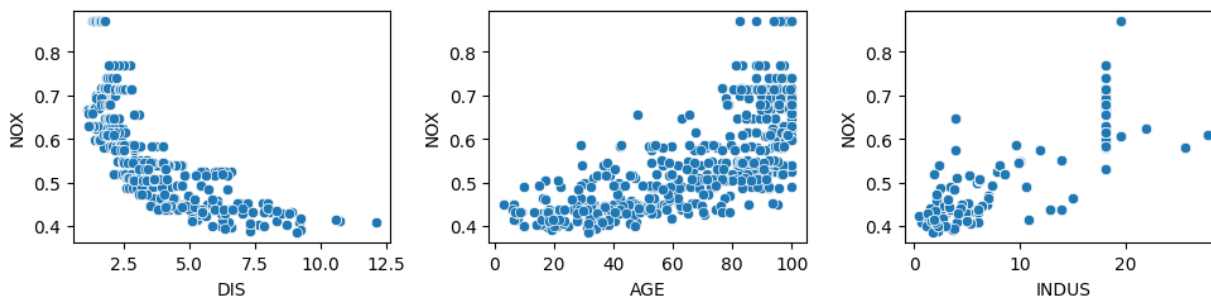
- There are 9 such relationships, represented with scatter plots below.

##### 1. Features with high correlations with TAX.



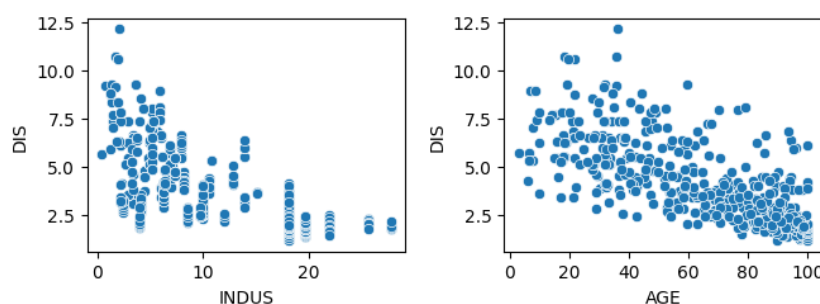
- TAX vs. RAD (0.91) – There is a high correlation but no visible trend. After removing the outliers, the correlation value drops to 0.25.
- TAX vs. INDUS (0.72) – Here some trend is visible, but it is not a strong one. Also, there are some outliers. After their removal correlation value drops to 0.23.
- The overall conclusion is that the tax rate for some properties might be higher due to other reasons.

##### 2. Features with high correlations with NOX.



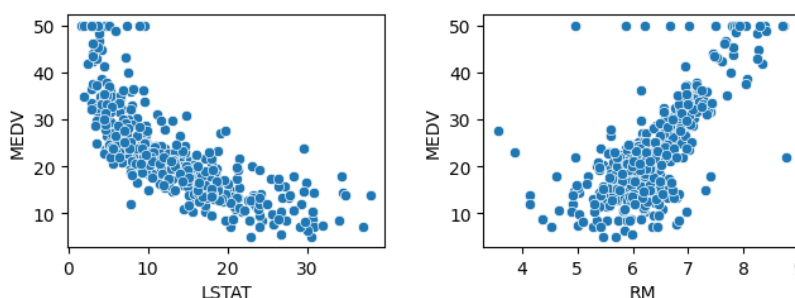
- NOX vs. DIS (-0.77) – There is a clear negative trend, suggesting that the five Boston employment centers are likely part of the industrial sector.
- NOX vs. INDUS (0.76) – A clear trend shows that as the proportion of non-retail business acres per town increases, so does air pollution. At an INDUS value of around 18, NOX levels rise, possibly due to additional factors or different types of industry.
- NOX vs. AGE (0.73) – A positive trend is visible, suggesting that some old houses may have been built close to industrial areas.

### 3. Features with high correlation with DIS.



- DIS vs. AGE (-0.75) – The distance of the houses to the Boston employment centers appears to decrease moderately as the proportion of old houses increase in the town. It is possible that the Boston employment centers are located in the established towns where proportion of owner-occupied units built prior to 1940 is comparatively high.
- DIS vs. INDUS (-0.71) – A clear negative trend is visible, most likely for the same reason as above.

### 4. Features with high correlation with MEDV.



- MEDV vs. LSTAT (-0.74) – The price of the house tends to decrease with an increase in LSTAT. This is also possible as the house price is lower in areas where lower status people live. There are also a few outliers, and the data seems to be capped at 50.
- MEDV vs. RM (0.7) – The price of the house seems to increase as the value of RM increases. This is expected as the price is generally higher for more rooms. There are also a few outliers in a horizontal line as the MEDV value seems to be capped at 50.

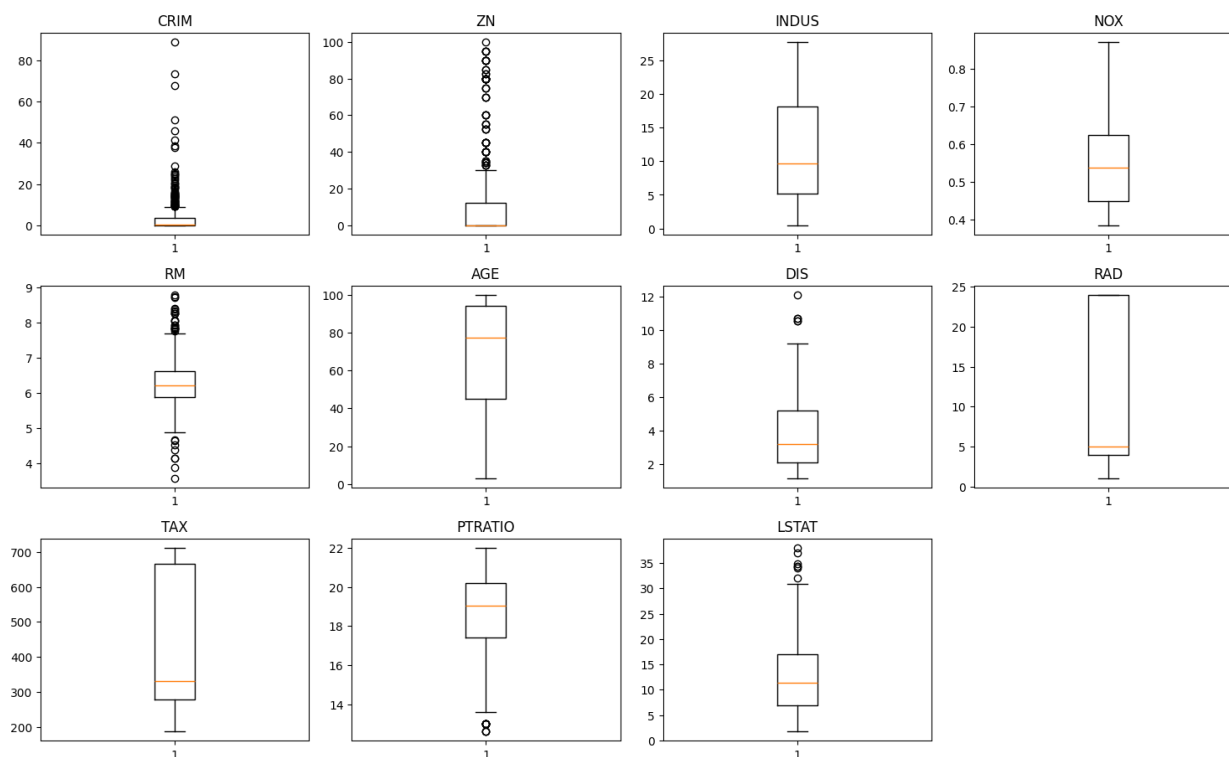
### 5. Bivariate analysis conclusions

- The analysis revealed some interesting relationships between features. However, some of these relationships are not desirable for linear regression models due to significant correlations between independent variables, such as NOX vs. INDUS, INDUS vs. AGE, and NOX vs. DIS.
- There are also 'hidden' relationships involving more than one factor, such as in the case of TAX.

- There is a linear relationship between the dependent variable MEDV and RM. After applying a log transformation to MEDV, this relationship will become non-linear, which may affect the quality of the linear regression models.
- Between MEDV and LSTAT, there is an  $e^{-x}$  relationship, indicating that a log transformation of MEDV will linearize this relationship.
- To address non-linear relationships between dependent and independent variables, I will apply non-linear models based on Decision Trees and compare the results.

## Missing Values and Outliers

- There were no missing values in the dataset.
- There are some outliers, but I have decided not to remove them as they do not appear to be incorrect data entries.
- In the case of crime statistics, these outliers may represent one-off effects. However, there is insufficient information about the data to justify their removal.
- Below figure illustrates residuals present in all numerical independent variables:



# Regression analysis

Before creating a linear regression model, I prepared the data and checked for multicollinearity.

## Data preparation

The data preparation stage was divided into two parts:

### 1. Split of the dataset

- Separation of dependent variable and independent variables.
- Split the dataset into training and testing sets in a 70/30 ratio.

### 2. Data scaling

- I have decided to scale the variables due to their diverse units and different scales, which should improve the understanding of the outcome parameters and their relationships.
- I have used `MinMaxScaler` due to the presence of dummy variable (CHAS),
- I have added a constant to the training and test datasets due to the use of the `statsmodels.api` library.

## Multicollinearity check

To check for multicollinearity, I used the Variance Inflation Factor (VIF). Features with a VIF score greater than 5 should be dropped or treated until all features had a VIF score below 5. The results are presented in the table below:

Feature	const	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT
VIF	89,26	1,92	2,74	4,00	1,08	4,40	1,86	3,15	4,36	8,35	10,19	1,94	2,86

- As expected, highly correlated values RAD and TAX (corr.: 0.91) had the VIF score above 5 indicating a strong relationship.

### NOTE:

After removing the outliers (which led to high correlation) from the TAX, only the NOX attribute had a VIF score above 5. However, as stated during the bivariate analysis, these outliers cannot be removed.

The next step was to remove the TAX column, as it had the highest VIF score, and check the multicollinearity again. The results are presented in the table below:

Feature	const	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	PTRATIO	LSTAT
VIF	89,26	1,92	2,48	3,27	1,05	4,36	1,86	3,15	4,33	2,94	1,91	2,86

- None of the VIF scores for the independent variables is above 5, indicating that multicollinearity has been removed.

## Metrics for models evaluation

To evaluate the performance of the models I used the following metrics:

- RMSE – root mean squared error,
- MSE – mean squared error,
- MAE – mean absolute error,
- MAPE – mean absolute percentage error,
- $R^2$  – the proportion of variance in the dependent variable explained by the independent variables.

## Building Linear Regression Model

The linear regression models were based on the Ordinary Least Squares (OLS) method.

### Model No.1

#### 1. Model evaluation

The metrics for the first model are presented in the table below:

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train	0.194892	0.037983	0.142894	0.049584	0.768629
Test	0.198258	0.039306	0.149346	0.051968	0.771996

- The error values for the train and test data are very close, with a tendency to be slightly higher for the test data set.
- The MAPE value is around 5% for both data sets, indicating good predictive accuracy.
- The  $R^2$  for the test data is slightly higher than for the training data. This might be caused by the presence of outliers – the ratio of outliers in the test set is higher than in the training set, therefore the model is slightly overfitting on the test data.
- To verify the final performance of the model a cross-validation will be applied.
- The  $R^2$  value shows that around 77% of the variance in the dependent variable (MEDV\_log) can be explained by the model.

#### 2. Examining the significance of the coefficients

The summary table for the model is shown below:

	const	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	PTRATIO	LSTAT
P> t	0	0	0,155	0,883	0,002	0	0,011	0,645	0	0	0	0

- INDUS, AGE, and ZN have p-values higher than 0.05, indicating that these features are not significant for predicting Y values (the null hypothesis cannot be rejected), and therefore, they can be eliminated.

## Model No.2

After eliminating the non-significant features from model no.1, I trained another model.

### 1. Model evaluation

The metrics for the second model are presented in the table below:

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train	0.195504	0.038222	0.143686	0.049818	0.767174
Test	0.198045	0.039222	0.151284	0.052580	0.772486

- The metrics show that after dropping three features, the second model does not perform worse than the first model.
- This means that I achieved the same performance with a simpler model.

### 2. Examining the significance of the coefficients

Just for a sanity check:

	const	CRIM	CHAS	NOX	RM	DIS	RAD	PTRATIO	LSTAT
P> t	0	0	0,002	0	0,0041	0	0	0	0

- The p-values for all the remaining features are less than 0.05, indicating that the parameters for the remaining features are significant (the null hypothesis can be rejected).
- The results also show that parameters of removed features had no predictive power.

## Linear Regression Assumptions

If the assumptions of the model are not satisfied, then the model might give false results. Hence, if any of the assumptions is not true, it is necessary to rebuild the model after fixing those issues.

### 1. Mean of residuals should be 0

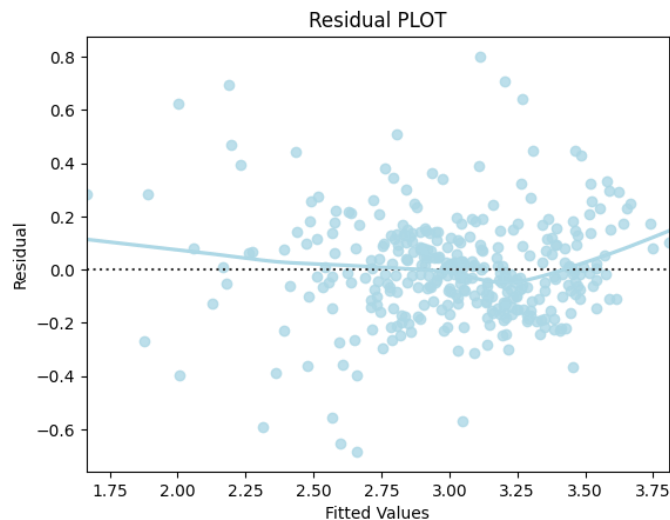
- The calculated mean of residuals equaled approximately  $4,99 * 10^{-15}$ .
- The value is very close to 0, therefore the assumption is satisfied.

## 2. No Heteroscedasticity

- To check this assumption, I used Goldfeldquandt Test with  $\alpha = 0.05$ .
- The p-value equaled **0.302** which is greater than 0.05 – the null hypothesis cannot be rejected.
- Residuals are homoscedastic, the assumption is satisfied.

## 3. Linearity of variables

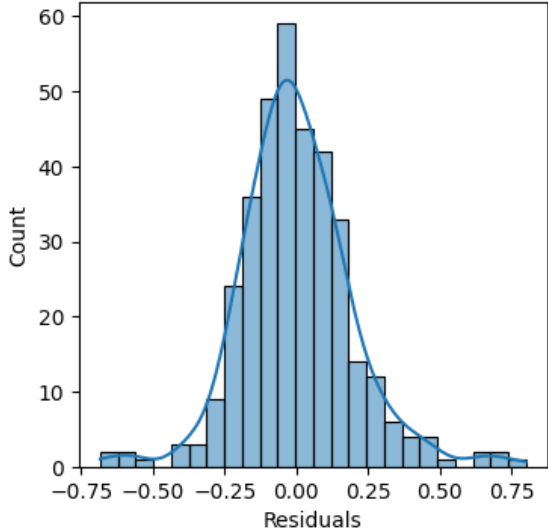
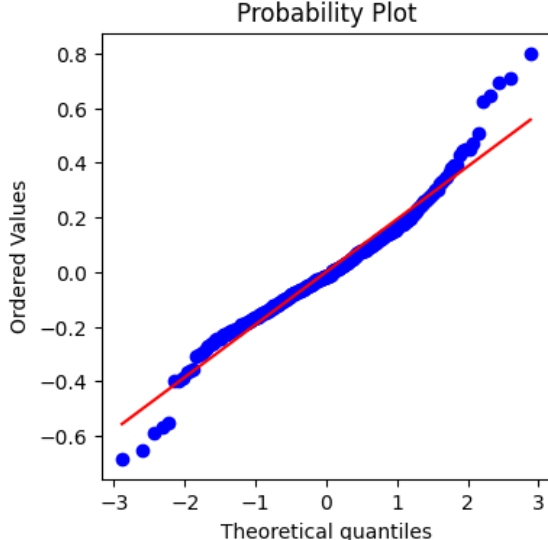
- To test the assumption, I plotted the relationship between the residuals and the fitted values



- There is no strong pattern in the residual vs. fitted values plot, the assumption is satisfied.
- Additionally, residuals are symmetrically distributed which confirms their homoscedasticity.

## 4. Normality of error terms

There are two approaches to verify this assumption:

Residuals distribution verification	QQ plot
	
<ul style="list-style-type: none"> <li>Residuals are normally distributed; <u>the assumption is satisfied</u>.</li> </ul>	<ul style="list-style-type: none"> <li>Residuals following normal distribution make a straight-line plot.</li> <li>QQ plot verifies residuals normal distribution, i.e., <u>the assumption is satisfied</u>.</li> </ul>

## 5. Summary

- All four assumptions of linear regression model are satisfied which means no additional modifications to the developed model are needed.

## Model Cross-Validation

To validate the developed model (Model No. 2), I employed a 10-fold cross-validation method, which is commonly used. The evaluation metrics for this model and the results of the cross-validation are presented in the table below:

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train	0.195504	0.038222	0.143686	0.049818	0.767174
Cross-Val.	0.200511	0.041093	0.149030	0.051735	0.729091

- The R<sup>2</sup> value for cross-validation is lower than for the training data set. This suggests that the developed model is overfitting.
- To reduce overfitting regularization methods should be considered.



## Regularization Methods

Given that cross-validation indicated that Model No. 2 was overfitting the data, I decided to apply regularization techniques to address this issue. Specifically, I used Ridge Regression, Lasso, and Elastic Net Regression to mitigate overfitting and compared their performance metrics.

For each regularization method, I performed a hyperparameter search using GridSearchCV to optimize the model's performance.

### 1. Ridge Regression

Before model tuning:

Model	Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Ridge	Train	0.197912	0.039169	0.141894	0.049283	0.761403
	Test	0.197211	0.038892	0.149990	0.051832	0.774398
Model No.2	Cross-Val.	0.200511	0.041093	0.149030	0.051735	0.729091

After model tuning ( $\alpha = 0.02$ ):

Model	Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Ridge Tuned	Train (cv)	0.195653	0.038280	0.142888	0.049571	0.766818
	Test	0.197088	0.038844	0.150507	0.052245	0.774681
Model No.2	Cross-Val.	0.200511	0.041093	0.149030	0.051735	0.729091

- Ridge Regression resulted in improved performance compared to the cross-validated results of Model No. 2, as evidenced by reduced error metrics and a higher R<sup>2</sup> value.
- Model tuning improved results, even though they are not significant.

### 2. Lasso

Before model tuning:

Model	Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Lasso	Train	0.405172	0.164164	0.297052	0.105047	0.000000
	Test	0.415866	0.172945	0.312503	0.107916	-0.003199
Model No.2	Cross-Val.	0.200511	0.041093	0.149030	0.051735	0.729091

After model tuning ( $\alpha = 0.001$ ):

Model	Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Lasso Tuned	Train (cv)	0.196841	0.038746	0.142682	0.049450	0.763978
	Test	0.197277	0.038918	0.150402	0.051938	0.774247
Model No.2	Cross-Val.	0.200511	0.041093	0.149030	0.051735	0.729091

- The error values of the tuned Lasso model are comparable with the cross-validated results of Model No.2
- The  $R^2$  value increased for the Lasso tuned model.

### 3. Elastic Net Regression

Before model tuning:

Model	Data	RMSE	MSE	MAE	MAPE	$R^2$
ElasticNet	Train	0.405172	0.164164	0.297052	0.105047	0.000000
	Test	0.415866	0.172945	0.312503	0.107916	-0.003199
Model No.2	Cross-Val.	0.200511	0.041093	0.149030	0.051735	0.729091

After model tuning ( $\alpha = 0.001$ ,  $L1\_ratio = 0.001$ ):

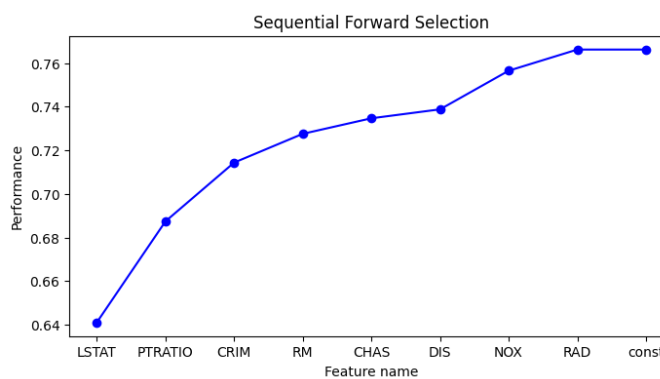
Model	Data	RMSE	MSE	MAE	MAPE	$R^2$
ElasticNet Tuned	Train (cv)	0.195929	0.038388	0.142428	0.049426	0.766161
	Test	0.196714	0.038696	0.150052	0.052040	0.775534
Model No.2	Cross-Val.	0.200511	0.041093	0.149030	0.051735	0.729091

- Since the Elastic Net Regression model combines both Lasso and Ridge Regression techniques, its error behavior is similar to the previous cases.
- The  $R^2$  value for the Elastic Net model on the test dataset is the highest among all the models evaluated.

### 4. Forward feature selection

In general, feature selection is used to reduce dimensionality and discard irrelevant or misleading features, resulting in faster training and testing processes. In my case, the model was already quite simple, with only nine features, and was not significantly overfitting. However, I aimed to identify the most relevant features to better understand their importance to the problem.

The Sequential Forward Selection algorithm was implemented on the Elastic Net Regression Tuned model.



## 5. Conclusions

- The differences in performance between the models with regularization methods applied are not statistically significant.
- The highest  $R^2$  value is achieved with Elastic Net Regression Tuned model.
- The lowest error values are observed by Ridge Regression Tuned model.
- Although regularization methods are intended prevent overfitting, all three models show slightly higher error values (RMSE, MAE, MAPE) and  $R^2$  parameter are slightly higher for the test data set than for the train set. This discrepancy may be due to the presence of outliers, which could be more prevalent in the test set.
- Looking at the SFS plot, the most significant feature is LSTAT, followed by PTRATIO and CRIM.

## Summary of the Linear Regression Models

- **Model Training and Evaluation:** During the analysis, I trained five models and evaluated their performance based on RMSE, MAPE, and  $R^2$  metrics.
- **Overfitting in Basic Linear Regression:** After applying cross-validation on the basic linear regression model (Model No.2), it appeared to be overfitting. Hence, I decided to use regularization methods such as Ridge Regression, Lasso, and Elastic Net Regression.
- **Impact of Regularization:** The trained models were not significantly different from each other, and all successfully reduced overfitting in the training model.
- **Performance Metrics:** The Mean Absolute Percentage Error (MAPE) was **5.2%**.
- **Best Model:** The Elastic Net Regression model explained around **77%** of the variation in the dependent variable, which was the highest among all the developed models.
- **Potential for Improvement:** While this result is decent, it could likely be improved with non-linear models, such as those based on Decision Trees.
- **Overfitting on Test Data:** The model still shows slight overfitting on the training data, so it is advisable to monitor and correct it if needed.
- **Model Equation:** The equation showing the relationship between the dependent variable and the independent variables is as follows:

$$\log(MEDV) = -0.81CRIM + 0.12CHAS - 0.47NOX + 0.33RM - 0.42DIS \\ + 0.15RAD - 0.44PTRATIO - 1.03LSTAT$$

## Models based on Decision Trees

- To find out if the model can be more predictive, i.e. better explain the dependent variable variations, I decided to create models based on Decision Trees.
- Such an approach does not need data scaling.

- I took into consideration several models from which I have chosen the best one.
- The evaluation metrics were the same as in the case of Linear Regression models – RMSE, MAPE,  $R^2$ .
- For each model I have applied a grid search method to find the best hyperparameters.

## 1. Decision Tree

Before tuning

Data	RMSE	MSE	MAE	MAPE	$R^2$
Train	6.675958e-17	4.456841e-33	1.003591e-17	2.868183e-18	1.000000
Test	0.1948107	0.03795122	0.1363801	0.04810739	0.779857

After tuning (max\_features = 1, max\_depth=5)

Data	RMSE	MSE	MAE	MAPE	$R^2$
Train (cv)	0.124842	0.015586	0.084638	0.029643	0.905061
Test	0.185271	0.034325	0.127413	0.045487	0.800889

- The model overfits training data.
- However, the performance on unseen data (test set) is good.

## 2. Random Forest

Before tuning

Data	RMSE	MSE	MAE	MAPE	$R^2$
Train	0.061825	0.003822	0.042834	0.015159	0.976716
Test	0.144290	0.020820	0.107639	0.038088	0.879232

After tuning (max\_features = 0.5, n\_estimators=120, max\_depth=None)

Data	RMSE	MSE	MAE	MAPE	$R^2$
Train (cv)	0.060935	0.003713	0.041271	0.014545	0.977382
Test	0.143390	0.020561	0.107626	0.038113	0.880734

- The model overfits training data.
- However, the performance on unseen data (test set) is very good.
- MAPE is less than 4%.

### 3. Ada Boost

Before tuning

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train	0.147975	0.021897	0.120439	0.041339	0.866617
Test	0.188561	0.035555	0.153560	0.052560	0.793755

After tuning (n\_estimators=110)

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train (cv)	0.149170	0.022252	0.124308	0.042643	0.864455
Test	0.189831	0.036036	0.155137	0.052960	0.790967

- The model overfits training data.
- The performance of the model decreased compared to previous models.
- The tuned model is slightly worse than the original one.

### 4. Gradient Boosting Regressor

Before tuning

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train	0.067177	0.004513	0.052134	0.017971	0.972511
Test	0.152809	0.023351	0.111362	0.039584	0.864550

After tuning (max\_features = 0.7, n\_estimators=120, max\_depth=5)

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train (cv)	0.022790	0.000519	0.017695	0.005997	0.996836
Test	0.149555	0.022367	0.109906	0.038860	0.870259

- The model overfits training data.
- However, the performance on unseen data (test set) is very good, comparable to Random Forest model.

## 5. XGBoost Regressor

Before tuning

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train	0.001590	0.000003	0.001138	0.000380	0.999985
Test	0.154053	0.023732	0.108923	0.038714	0.862337

After tuning (n\_estimators=100, max\_depth=None)

Data	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Train (cv)	0.001590	0.000003	0.001138	0.000380	0.999985
Test	0.154053	0.023732	0.108923	0.038714	0.862337

- The model overfits training data.
- However, the performance on unseen data (test set) is good, but worse compared to Random Forest model.
- Tuning did not bring improvements to the model.

## 6. Performance summary of tuned models

Type	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Decision tree	0.185271	0.034325	0.127413	0.045487	0.800889
Random Forest	0.143390	0.020561	0.107626	0.038113	0.880734
Ada Boost	0.188561	0.035555	0.153560	0.052560	0.793755
Gradient Boosting	0.149555	0.022367	0.109906	0.038860	0.870259
XG Boost	0.154053	0.023732	0.108923	0.038714	0.862337

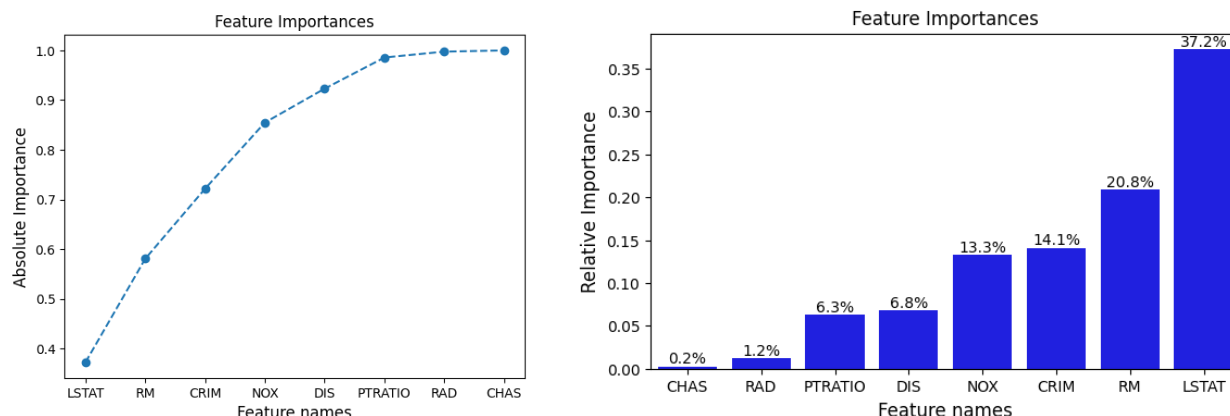
- According to metrics, the Random Forest model achieves the best performance among the models evaluated.
- The MAPE of **3.8%** shows a significant improvement compared to the Linear Regression model, with a **~27%** enhancement – predictions are more accurate.
- The Random Forest model exhibits some overfitting, performing extremely well on training data but still performing very well on the test data.
- The R<sup>2</sup> of **88%** on test data indicates that the model explains a significant portion of the variance, despite overfitting concerns.
- The Random Forest model represents a substantial improvement over the Linear Regression model, with a **~13%** better R<sup>2</sup>.

Type	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Random Forest	0.143390	0.020561	0.107626	0.038113	0.880734
Elastic Net Tuned	0.196714	0.038696	0.150052	0.052040	0.775534

- The overfitting of the Random Forest Tuned model can be further targeted by exploring a wider range of hyperparameters, e.g. `min_samples_split`, `min_samples_leaf`, or `bootstrap`, or by decreasing the maximum depth for the trees.

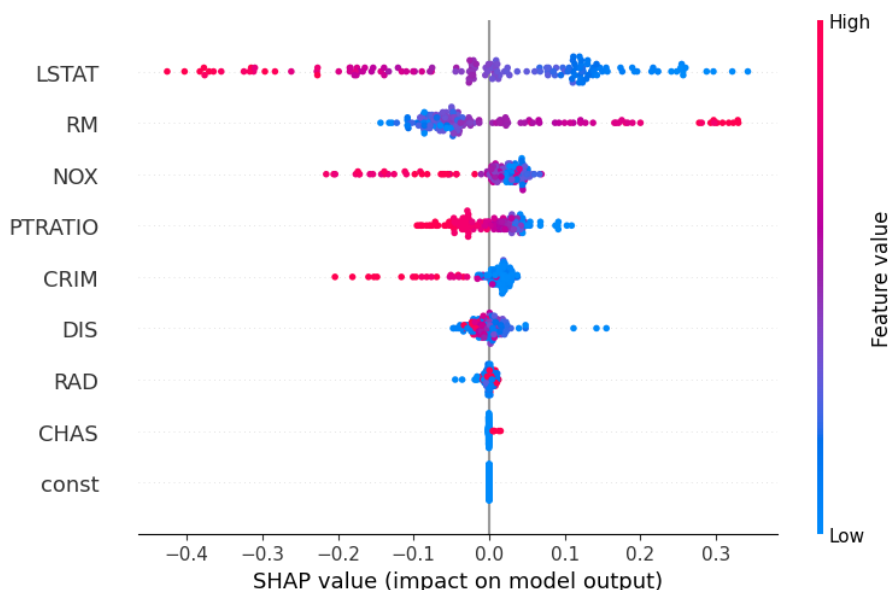
## Features Importance

Similar to the Linear Regression model, I examined the feature importance in the Random Forest model to better understand the relationships between variables.



- The bar plot indicates that the most important features in the model are LSTAT, RM, CRIM, and NOX. These features have the highest influence on the model's predictions.
- The feature representing boundaries with the river (CHAS) has a nearly negligible impact on the model's predictions, suggesting it is not a significant factor in this context.

To see the direction of the relationships between features I used SHAP Values. The results are presented below:



- The plot indicates that LSTAT, NOX, CRIM, and PTRATIO have a negative impact on the predictions.
- RM shows the most positive relationship with the predicted values.
- DIS and RAD have a moderately neutral influence.
- CHAS has a negligible effect on the model's predictions.

## Summary of the Decision Tree Models

- **Model Training and Evaluation:** During the analysis, I trained five models and evaluated their performance based on RMSE, MAPE, and  $R^2$  metrics.
- **Hyperparameter Tuning:** Each model was tuned and cross-validated using GridSearchCV.
- **Model Diversity and Overfitting:** The trained models showed noticeable diversity; however, all exhibited overfitting on the training data.
- **Best Performance:** The best Mean Absolute Percentage Error (MAPE) was achieved by the Random Forest model, with a value of **3.8%**.
- **Explained Variance:** The Random Forest model was able to explain **88%** of the variation in the dependent variable, the highest among all the developed models, including Linear Regression.
- **Overfitting and Accuracy:** While the Random Forest model overfits the training data, it provides very accurate predictions on the test data. Therefore, it is advisable to monitor this model and apply corrections if necessary.

## Insights for the business

### 1. Linear Regression Model

- **Linear Regression Equation:**

$$\log(MEDV) = -0.81CRIM + 0.12CHAS - 0.47NOX + 0.33RM - 0.42DIS + 0.15RAD - 0.44PTRATIO - 1.03LSTAT$$

- **Performance of the model:**

Type	RMSE	MSE	MAE	MAPE	$R^2$
Elastic Net Tuned	0.196714	0.038696	0.150052	0.052040	0.775534

## Key Insights

- **Scaled Features:** Features are scaled in the range 0 – 1, meaning a one-unit change in the variable refers to a change by its maximum value.



- **Log Scale Transformation:** MEDV is in the log scale, so final predictions must be transformed using  $e^x$ .
- **Percentage Change Interpretation:** Due to the log scale, the coefficients represent the percentage change in the output value with a one-unit change in a particular variable, keeping other factors constant. Specifically:

$$MEDV\% = e^{\theta} - 1 * 100\%$$

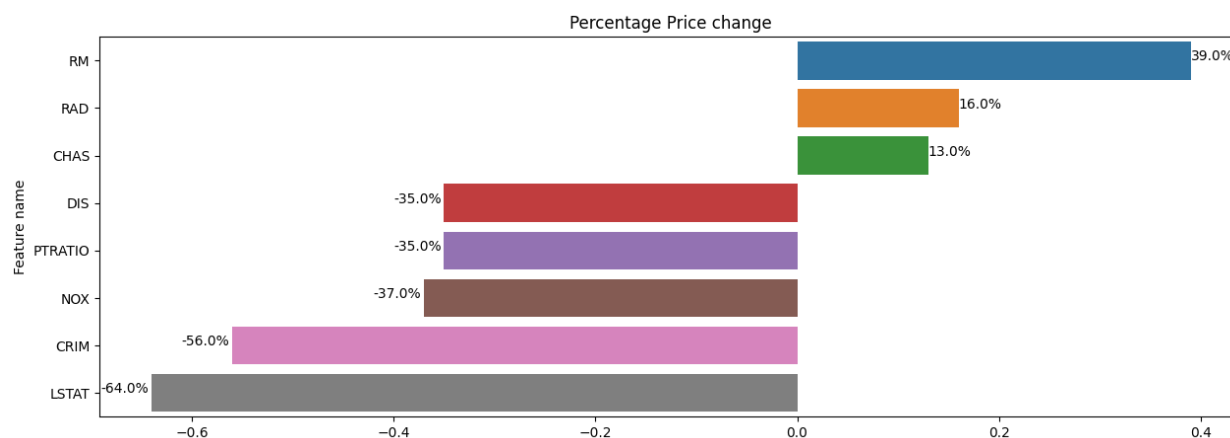
Where:

$\theta$  – independent variable coefficient.

- **Feature Ranges before Normalization**

	Min	Max	Range
CRIM	0.00632	88.9762	88.97
CHAS	0	1	1.00
NOX	0.385	0.871	0.49
RM	3.561	8.78	5.22
DIS	1.1296	12.1265	11.00
RAD	1	24	23.00
PTRATIO	12.6	22	9.40
LSTAT	1.73	37.97	36.24

- **Visualization:** The chart below illustrates the percentage influence of variables on the MEDV value:



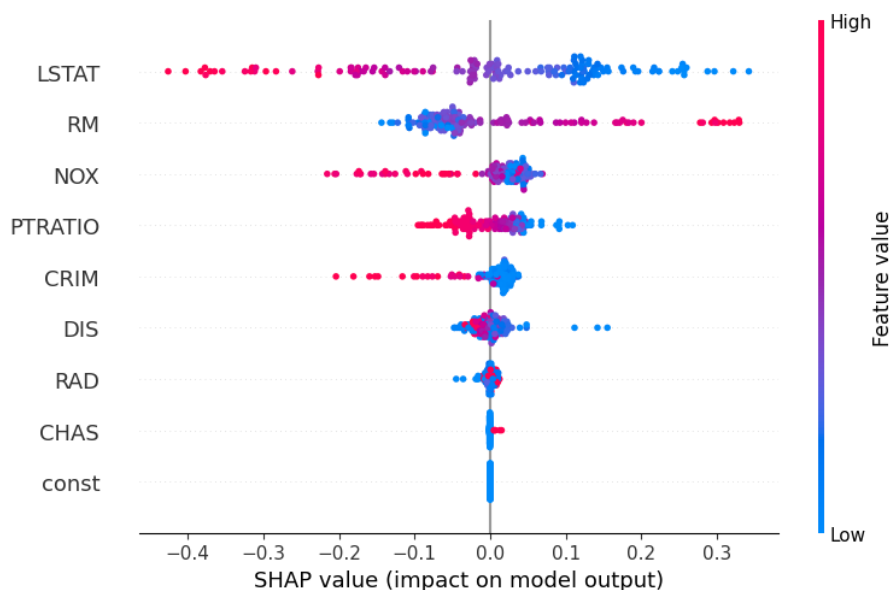
## 2. Random Forest Regression Model

- **Performance of the model:**

Type	RMSE	MSE	MAE	MAPE	R <sup>2</sup>
Random Forest	0.143390	0.020561	0.107626	0.038113	0.880734

## Key Insights

- **Visualization:** The chart below illustrates features importance and their impact on the output model.



### 3. Key insights for decision making:

#### 1. Target High-Impact Features:

- **LSTAT (Percentage of Lower Status Population):** This feature has the highest negative impact on MEDV. Reducing the percentage of lower status residents through community programs, education, and economic development could significantly increase property values.
- **RM (Average Number of Rooms per Dwelling):** This has the highest positive impact. Encouraging or facilitating the construction of homes with more rooms could boost property values.
- **NOX (Nitric Oxides Concentration):** This has a substantial negative impact. Efforts to reduce pollution and improve air quality could positively influence property values.
- **CRIM (Per Capita Crime Rate):** This negatively impacts property values. Implementing effective crime reduction strategies could lead to higher property values.

#### 2. Monitor Moderate-Impact Features:

- **DIS (Weighted Distance to Employment Centers):** While the impact is moderate and negative, improving transportation infrastructure to reduce travel time to employment hubs could have a beneficial effect.
- **PTRATIO (Pupil-Teacher Ratio):** This also has a moderate negative impact. Enhancing educational resources and reducing the pupil-teacher ratio in schools could improve property values.

- **RAD (Index of Accessibility to Radial Highways):** This feature has a moderate impact and is slightly positive. Ensuring good accessibility while balancing other factors like noise pollution could be beneficial.

### 3. Negligible Impact Features:

- **CHAS (Proximity to Charles River):** This feature has a negligible impact on property values. Investments related to the proximity to the river may not yield significant returns in terms of property value increase.

# Conclusion

The analysis led to selecting the Random Forest Regression model due to its superior performance on the training dataset compared to the Linear Regression model. Based on the insights from this model, the business can take the following actions:

## 1. **Community:**

- Implement socio-economic improvement programs targeting lower-status populations to enhance their living conditions.
- Invest in local businesses, create job opportunities and indirectly increase property values.

## 2. **Housing Development:**

- Promote the construction of larger homes with more rooms to attract higher home values by attracting higher-income residents.
- Balance development by including affordable housing to ensure a diverse and inclusive community.

## 3. **Environmental Policies:**

- Work with local government to strengthen environmental regulations to reduce NOX emissions and improve air quality contributing to higher property values.
- Promote green energy initiatives.
- Increase the number of parks and green spaces to improve air quality and make neighborhoods more attractive.

## 4. **Educational Investments:**

- Advocate for increased funding for schools to lower the pupil-teacher ratio, thereby making the area more attractive for families and increasing property values.
- Support local schools with extracurricular programs and tutoring services to enhance educational outcomes.

## 5. **Crime Reduction:**

- Develop comprehensive crime reduction strategies to improve community safety and attractiveness.
- Create environments that discourage criminal activity through urban planning, invest in security systems, improve street lighting
- Implementing programs aimed at reducing crime rates.

## 6. **Transportation and Infrastructure:**

- Invest in transportation infrastructure to reduce commute times to employment centers and improve accessibility to highways, enhancing the desirability of the area.

- Invest in better road maintenance and expanding public transport routes.

**7. Resource Allocation:**

- Focus resources on impactful areas such as socio-economic improvement, housing development, education, and environmental quality rather than on proximity to the river, which has a negligible impact.
- On the other hand, it is important to recognize the potential benefits offered by natural surroundings. Promoting the development of recreational and commercial amenities near the Charles River can be beneficial for enhancing the attractiveness of nearby properties.