**Linear Regression on The Boston Housing Dataset using sklearn**

**List of Contents**

***Objective Scope***

***Data Description***

***Data Preview***

***Data Preprocessing and Cleaning***

***Exploratory Data Analysis (EDA)***

***Univariate & Bivariate analysis and their Inference***

***Outlier Treatment***

***Train and Test Split***

***Training the Linear Regression Model***

***Prediction***

***Evaluating Model performance***

***Model Summary***

***Conclusion***

***Suggestions for Home buyers***

***Author’s Thoughts***

**Code** : <https://github.com/PrasannaWorkshop/Regression/blob/main/Regression_Boston_Housing.ipynb>

**The Boston Housing Dataset**

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of [Boston MA](http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html).

**Objective**

The goal of this exercise is to predict median house values using linear regression in scikit-learn. This involves using a dataset, like the Boston Housing Dataset, to train a linear regression model that can estimate house prices based on various features.

The model is then evaluated using metrics like R2 score to assess its predictive accuracy.

**Scope**

Although there are various regression techniques such as Linear Regression, Logistic Regression, Polynomial Regression, Ridge Regression, Lasso Regression, and Decision Tree Regression,**The scope of this exercise is limited to demonstrating Exploratory Data Analysis (EDA) and Linear Regression.**

*Separate blogs or exercises covering the other regression techniques will follow.*

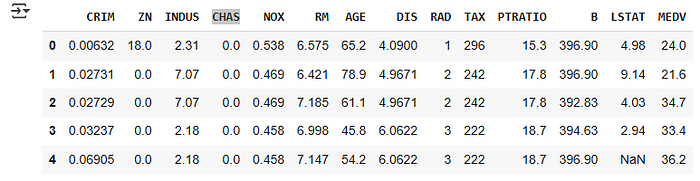
**Data Description**

The following describes the dataset columns

**CRIM**: Per capita crime rate by town.  
**ZN:** Proportion of residential land zoned for large 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 oxides concentration (parts per 10 million).  
**RM:** Average number of rooms per dwelling.  
**AGE:** Proportion of owner-occupied units built prior to 1940.  
**DIS:** Weighted distances to five Boston employment centers.  
**RAD:** Index of accessibility to radial highways.  
**TAX:** Full-value property-tax rate per $10,000.  
**PTRATIO** — pupil-teacher ratio by town  
**B**: 1000(Bk — 0.63)² where Bk is the proportion of blacks by town  
**LSTAT**— % lower status of the population  
**MEDV**— Median value of owner-occupied homes in $1000's

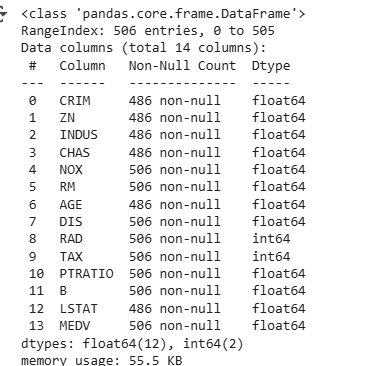
**Data Preview**

**Lets check the first 5 rows**



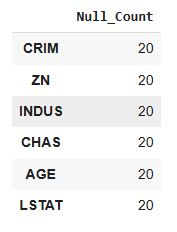
Data seems to be loaded properly

**Lets check the column information**



The dataset contains 506 rows and 14 columns, of which 12 columns are of float type and 2 columns are of integer type.

We observe that some columns have a non-null count less than 506, indicating missing values in the column

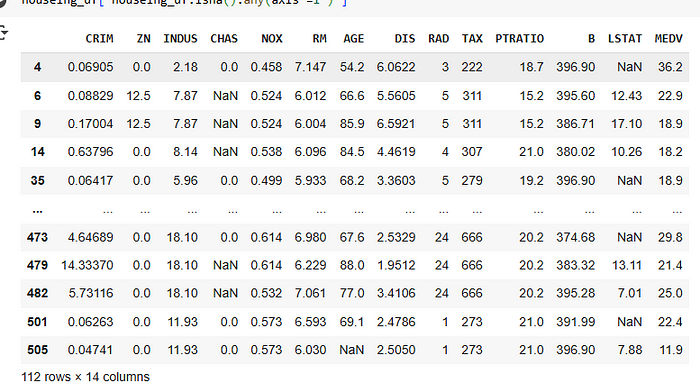


**Data Preprocessing and Cleaning**

**Treating the null Values**

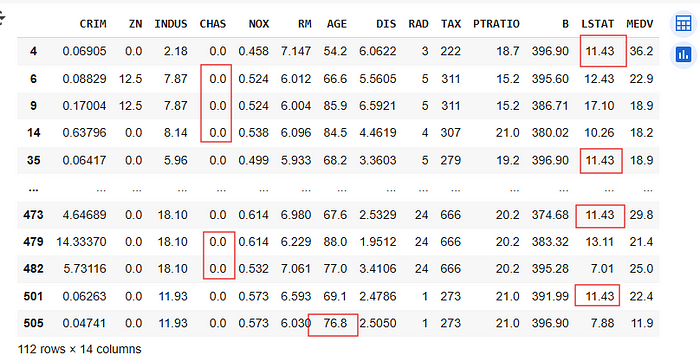
We observed that some columns contain null values. To address this, we will replace the missing values with the median of the respective columns.

There are 112 rows with nulls present in one or more columns



We chose to replace missing values with the median for each column, as most of the columns are skewed and the mean may not be a reliable measure of central tendency

After replacing the nulls with median for each column

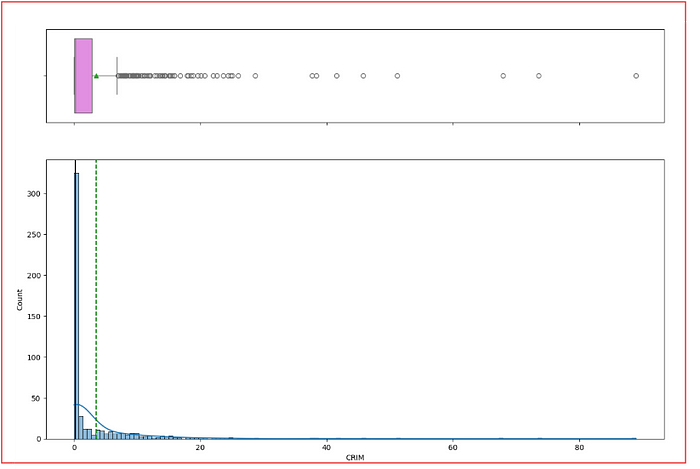


**Exploratory Data Analysis (EDA)**

**Univariate data analysis**

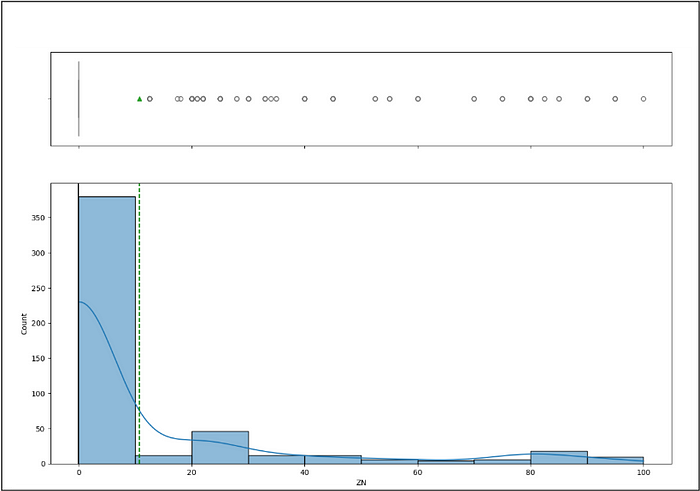
*Univariate analysis is a statistical method that focuses on a single variable within a dataset. Its primary goal is to describe and summarize the characteristics of that variable, such as its central tendency (mean, median, mode) and variability (range, standard deviation). It helps in understanding the distribution of the variable and identifying patterns within it.*

**Univariate analysis of CRIM**



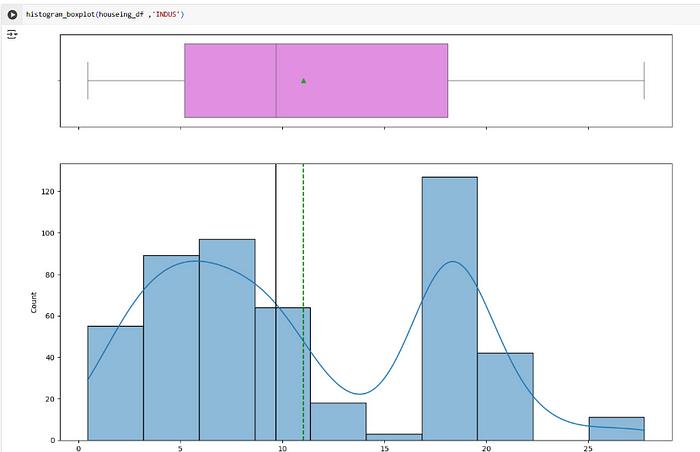
**## Inference**

CRIM :Crim ( Per capita rate by town ) is right skewed . The distribution shows that most of the area have low crime rate , with few area having higher crime rate values . These are outlier and needed to be treated

**Univariate analysis of ZN**

**## Inference**zn: Proportion of residential land zoned for large lots (over 25,000 sq. ft.)

The ZN is right skewed .Meaning most of the area having zn proption to zero indicating that there are very less land that is reserved for large lots over 25000sqft ,i.e. most of the land are of small lots

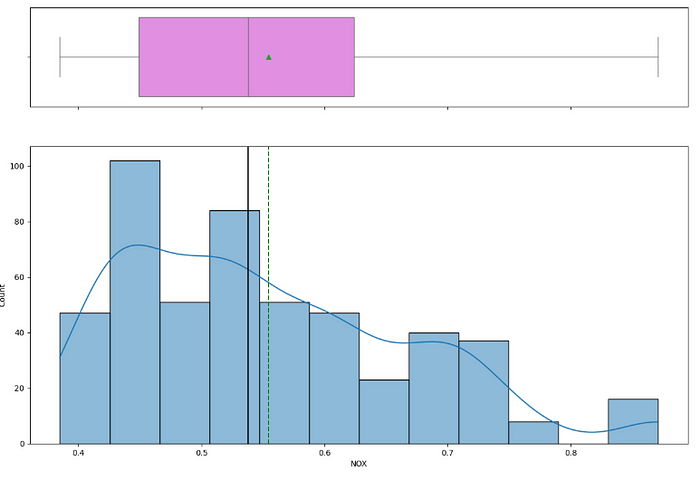
**Univariate analysis of INDUS** 

**## Inference :**Indus (Proportion of non-retail business acres per town.)

From the histogram we see that the Indus is bimodal ( 3–9 and 18–20)

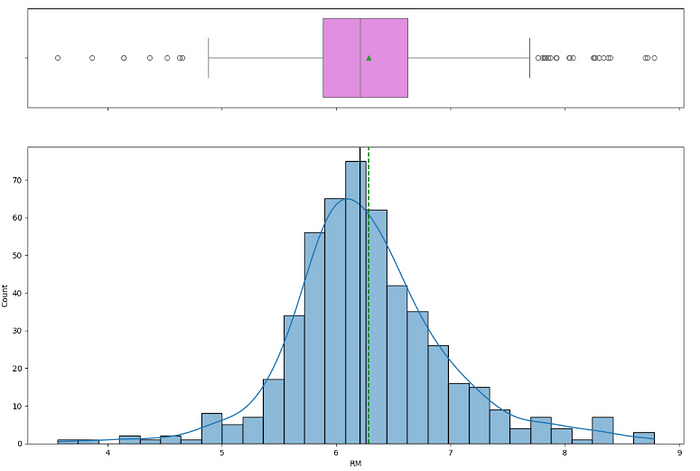
The distribution is complex which is not symmetric nor it is skewed

This means that few areas might have more portion of non-retail business or few might have less portion.

**Univariate analysis of Nox**

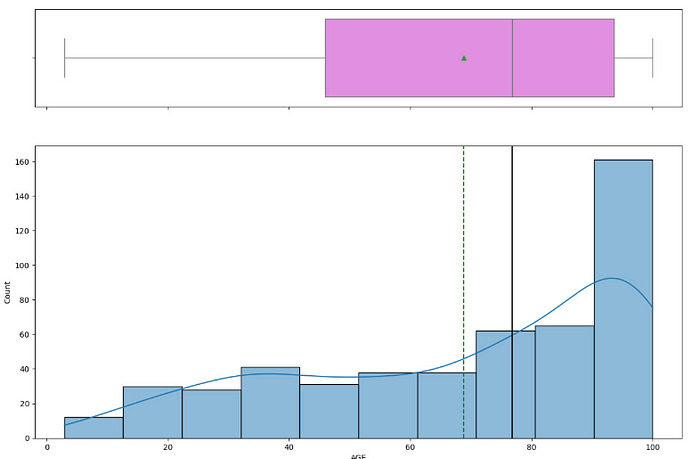
**## Inference**NOX: Nitric oxides concentration (parts per 10 million)

. The histogram or distribution is right skewed .With median and average between 0.5 to 0.6 indicating the Nox level are low in most of the areas

**Univariate Analysis of RM**

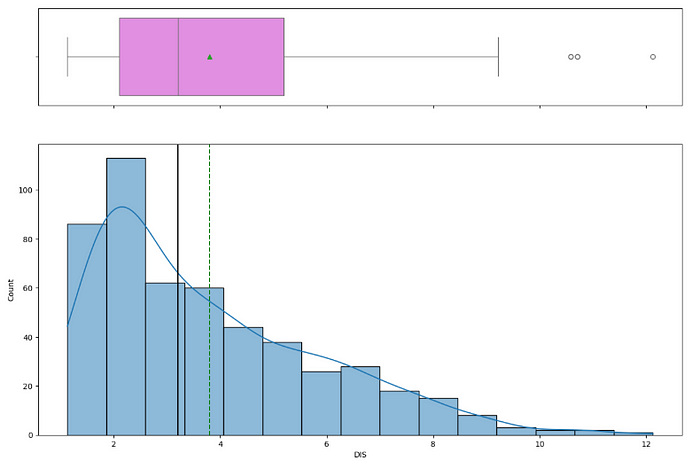
**## Inference** rm: Average number of rooms per dwelling.

Rm variable is almost normally distributed which is skewed slightly to the right Most of the areas have 6.2 to 6.3 rooms

**Univariate Analysis of Age**

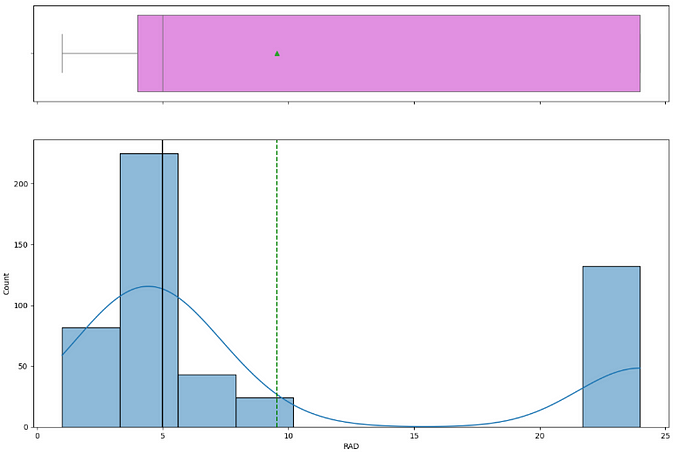
**## Inference**age: Proportion of owner-occupied units built prior to 1940.

The data is skewed towards left having max count at 100 , indicating many of the houses are quit old .

**Univariate Analysis of DIS**

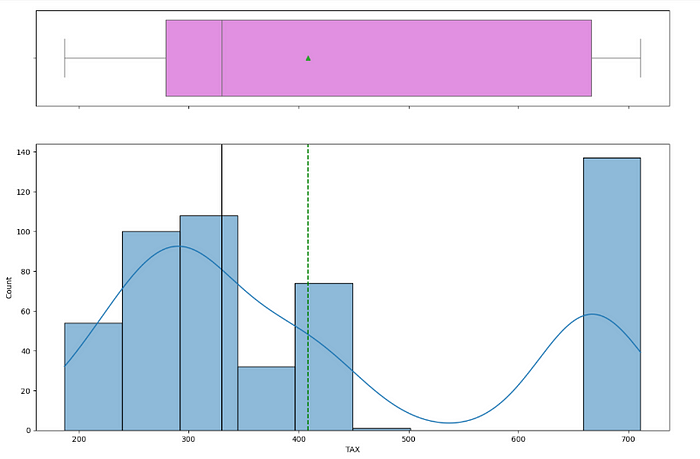
**## Inference :**dis — Weighted distances to five Boston employment centers

The distribution is skewed right and mode of the distribution is 2–3 ,meaning most of the houses are closer to the employment center

**Univariate Analysis of RAD**

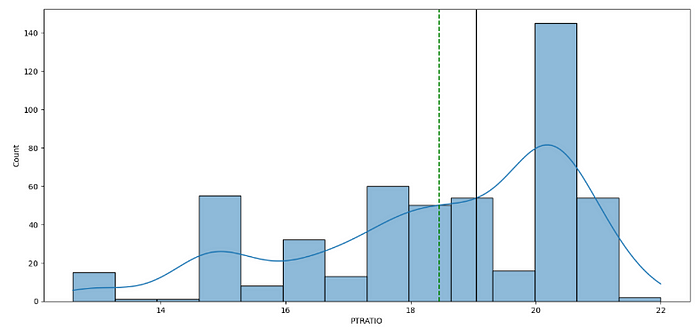
**## Inference rad:** Index of accessibility to radial highways.

we see a complicated distribution it has two modes 5 and 22–24 meaning there are two cluster of houses one who are located near to highways another cluster who are far away from the radial highways .

**Univariate Analysis of Tax**

**Inference:** tax: Full-value property-tax rate per $10,000.

we see a complex distribution with two peaks ie , meaning two cluster one having are being low tax paying area and one cluster payer high tax paying area

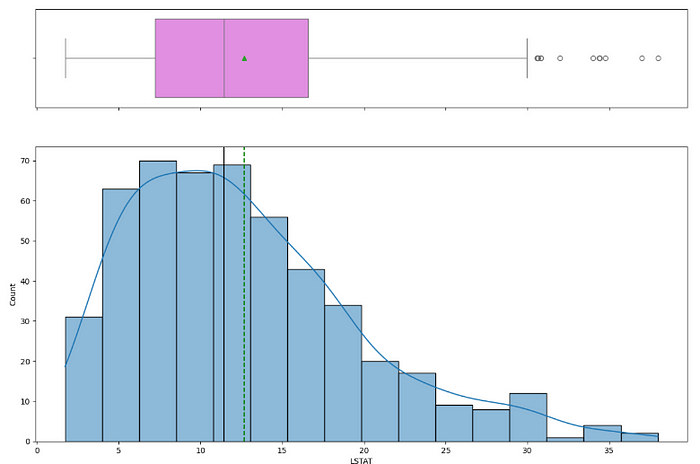
**Univariate Analysis of PTRATIO**

**## Inference : PTRATIO — pupil-teacher ratio by town**

The distribution is left skewed

max PTRATIO is 20 meaning most of the areas have very less teacher than students . meaning their are more children in the area or schools have large student population or more number of students per classroom

**Univariate Analysis of LSTAT**

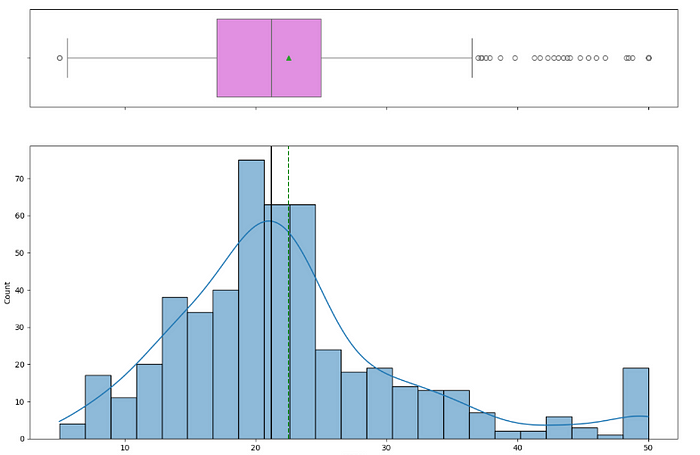


**## Inference** :LSTAT — % lower status of the population

*“Lower status of the population” generally refers to individuals or groups with less social and economic power, influence, and opportunities compared to those with higher status*

The LSTAT distribution is right skewed . Majority of the count is between 5–15 and these are Low percentage LSTAT area . Hence most of the area population are having low social and economic condition

**Univariate Analysis of MEDV**



**## Inference** MEDV — Median value of owner-occupied homes in $1000's

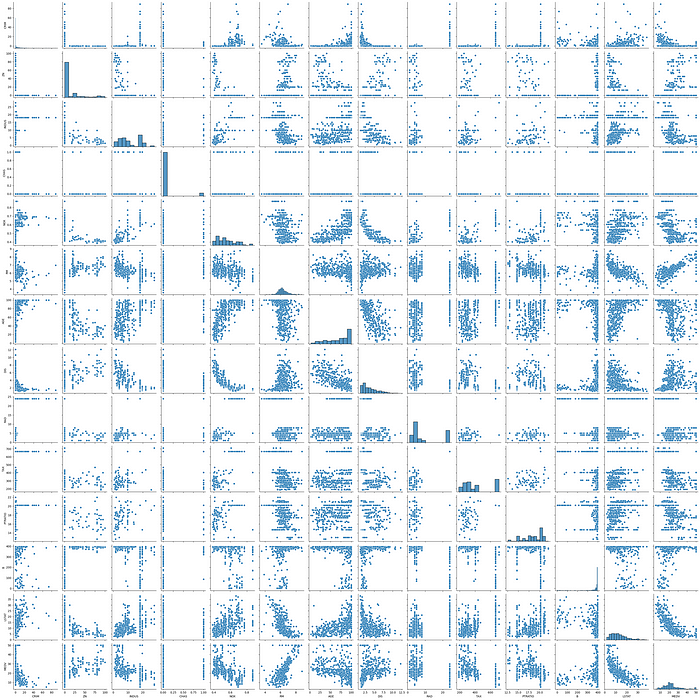
The distribution is almost normal and is right skewed .Most of the houses are between `$19k -$24k range .There are few outlier houses more than ‘50k dollar’ are capped at 50k dollar

\* **This is our target varaible / dependent variable/ Response variable**

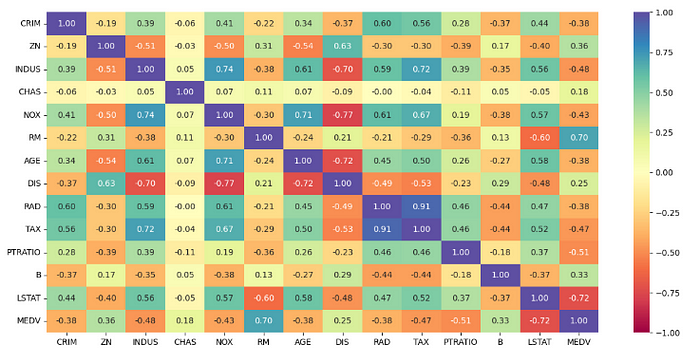
**Bivariate Analysis**

*Bivariate analysis is a statistical method that investigates the relationship between two variables. It examines how changes in one variable are associated with changes in another. This analysis helps determine if there’s a statistical association, the strength of that association, and whether one variable can be predicted from the other.*

**Pair plot**



**Correlation Matrix**



**## Inference**

\* MEDV : few columns like RM and LSTAT have strong correlation with MDEV ,NOx , PTRATIO also have good correlation with MDEV which would mean this predictors will play a vital role in predicting the the Median home value

\* Few of the predictors or independent variables are strongly correlated to each other , which indicates multicollinearity is present , hence we might have to drop few independent variables or combine them to create a new one

**Outlier Treatment**

The data has outlier in most of the columns

*Detecting Outlier with IQR*

*IQR is used to****measure variability****by dividing a data set into quartiles. The data is sorted in ascending order and then we split it into 4 equal parts. The values Q1 (25th percentile), Q2 (50th percentile or median) and Q3 (75th percentile) separate dataset in 4 equal parts.*

*If a dataset has*2n or 2n+1*data points, then*

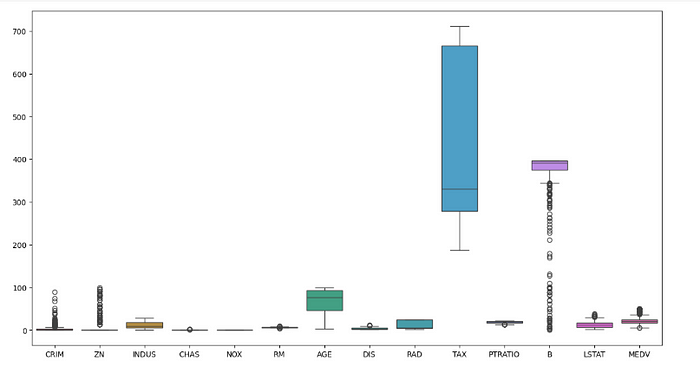
*Q2 = median of the dataset.*

*Q1 = median of n smallest data points.*

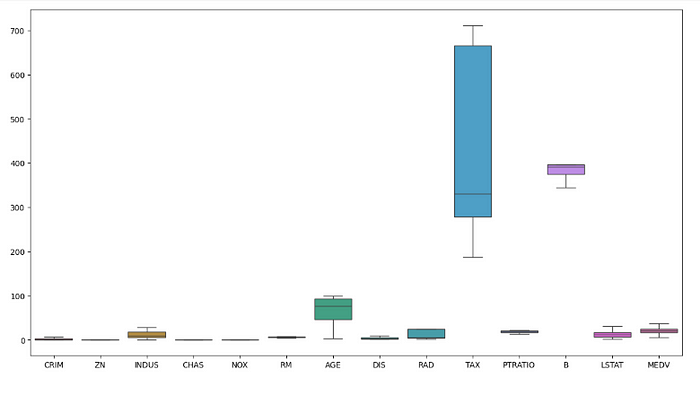
*Q3 = median of n highest data points.*

*The IQR is calculated as:****IQR=Q3−Q1 = Q3 — Q1***

***Data points that fall below Q1−1.5×IQR or above Q3+1.5×IQR are considered outliers***



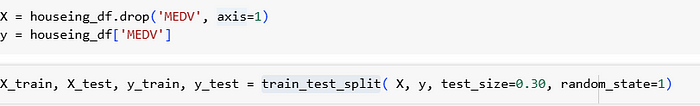
**After Outlier Treatment**



**Conclusion** : all the outliers from every column are removed

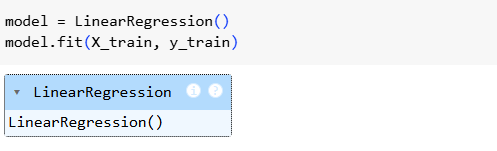
**Train and Test Split**

*The****train\_test\_split()****method from****sklearn.model****is used to split our data into train and test sets. First, we need to divide our data into features (X) and labels (y). The dataframe gets divided into X\_train, X\_test, y\_train, and y\_test. X\_train and y\_train sets are used for training and fitting the model. The X\_test and y\_test sets are used for testing the model if it’s predicting the right outputs/labels.*



**Training the Linear Regression Model**

*Model training is the machine learning (ML) step where the “learning” occurs. In machine learning, learning involves adjusting the parameters of an ML model. These parameters include the*[*weights and biases*](https://www.ibm.com/think/topics/backpropagation#How+neural+networks+work)*in the mathematical functions that make up their algorithms. The goal of this adjustment is to produce more accurate outputs. The specific values for these weights and biases, which are the end result of model training, are the tangible manifestation of a model’s “knowledge.”*

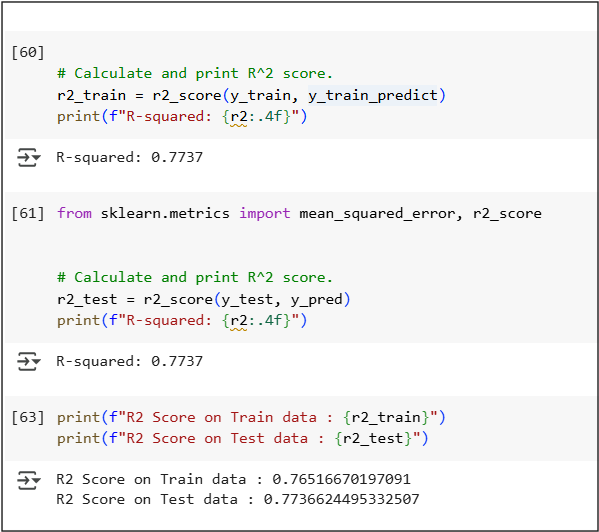


**Prediction**

We use X\_test to generate predictions stored in y\_pred, which contains the outputs predicted by the model. These predictions are then compared with the actual values in y\_test to evaluate the model's performance and determine its accuracy.



**Evaluating Model performance**

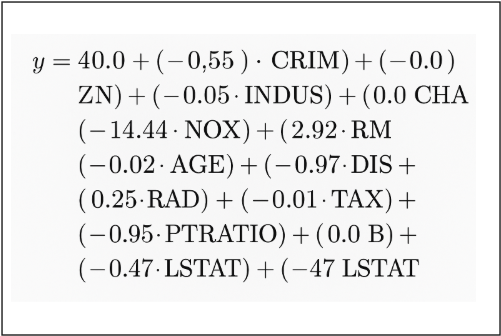


Inference

* The model performs well with good generalization.
* **R² on training data:** 0.765 → explains ~76.5% of the variance.
* **R² on test data:** 0.774 → explains ~77.4% of the variance on unseen data.

Since the test score is slightly higher than the training score, the model is **not overfitting** and is likely well-tuned.

Regression Equation:



**Model Summary**

The given linear regression equation predicts a target variable yyy (likely house price) using multiple predictors. Here’s a summary of the key influences:

* **Positive impact on yyy** (increases the value):
* **RM (avg. rooms per dwelling):** +2.92 → strong positive effect.
* **RAD (accessibility to highways):** +0.25 → small positive effect.
* **Negative impact on yyy** (decreases the value):
* **NOX (nitric oxide concentration):** −14.44 → strong negative effect.
* **PTRATIO (pupil-teacher ratio):** −0.95
* **DIS (distance to employment centers):** −0.97
* **LSTAT (% lower status population):** −0.47
* **CRIM (crime rate):** −0.55
* **INDUS (industrial proportion):** −0.05
* **AGE, TAX:** small negative effects
* **No effect:** Coefficients of 0 for **ZN, CHAS, B** → these variables do not influence yyy in this model.

**Conclusion:**

* The number of rooms (**RM**) boosts house value significantly.
* High pollution (**NOX**) and socioeconomic indicators like **LSTAT** and **CRIM** reduce it.
* Some variables are statistically insignificant in this model.

**Suggestions for Home buyers**

(RM) : Aim for properties with **more rooms for better value and resale**  
(DIS) : Distance from employment , **look for areas that are near to Employment area**  
(PTRATIO) : Lower pupil to teacher ratio indicates better school , which drive the house demand , **Look for schools with good education quality**

(CRIM) : Crime rate effects the home value negatively , **avoid the areas with high crime rate**  
(NOXO : **Avoid the areas with high Nitrogen oxide concentration** , which reduces the property value  
LSTAT : **Look for the areas with low LSTAT va**lues (% of lower status population)

**Author’s Thoughts:**

1. If you like the post, please like the post
2. There are many advance techniques such as (Lasso and ridge) to increase the accuracy of the model that will be posted in separate post
3. Another post about the CLRM (Classical Linear regression model) will also be posted