**Boston Housing Data Project Write-up(Short)**

Introduction:

The project explores the Boston dataset with Linear Modelling algorithm, to develop a model to predict prices of Houses in Boston, analyzing different variables in dataset.

Source of Data:

The data for the project is taken from UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/housing

This data is also available packages with sklearn package, a standardized dataset was utilized to iron-out probable issue with inconsistency of data.

The data contains about 500 instances and 14 attributes. The sample values below columns:

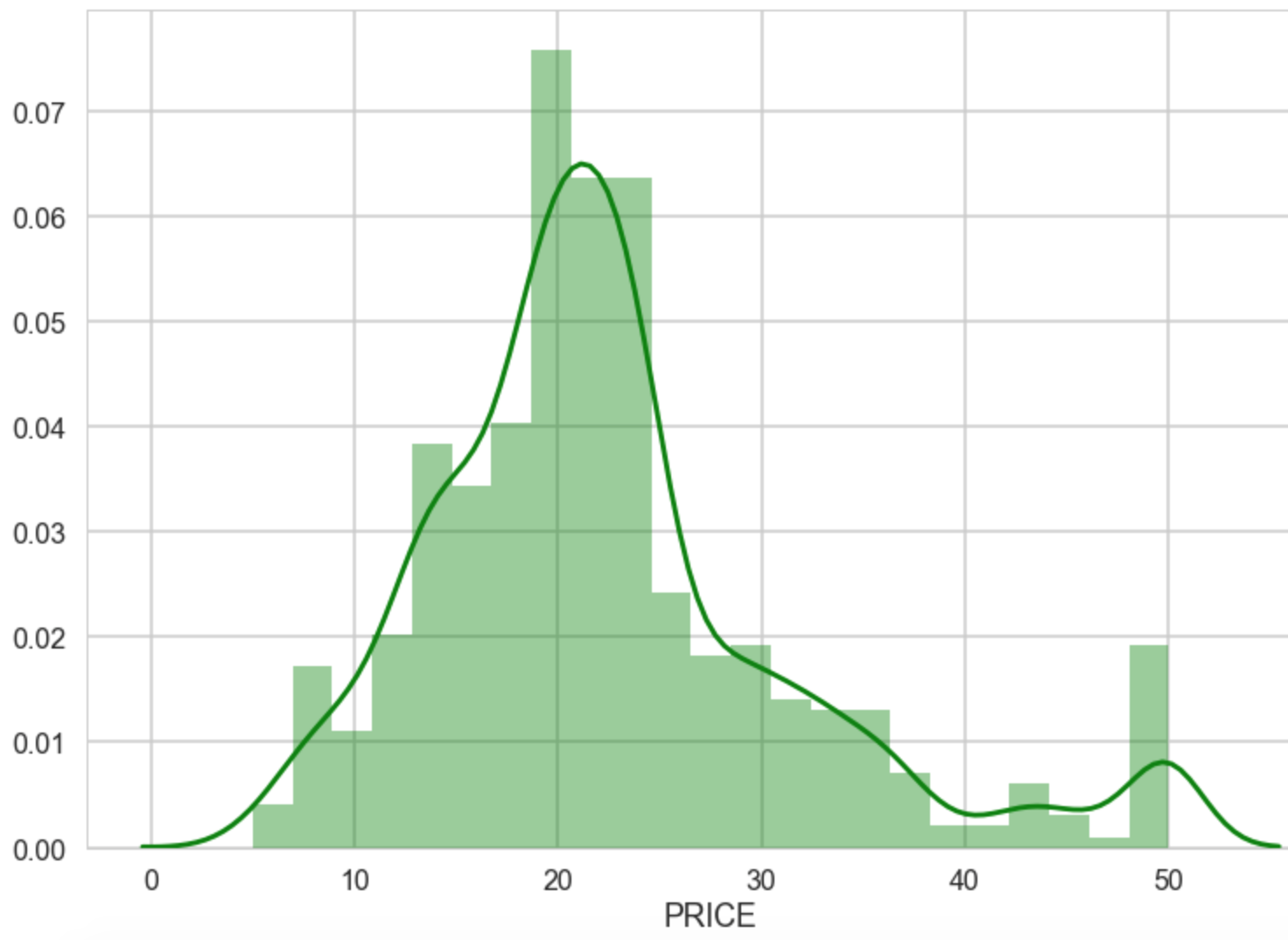
1. CRIM: per capita crime rate by town   
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.   
3. INDUS: proportion of non-retail business acres per town   
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)   
5. NOX: nitric oxides concentration (parts per 10 million)   
6. RM: average number of rooms per dwelling   
7. AGE: proportion of owner-occupied units built prior to 1940   
8. DIS: weighted distances to five Boston employment centres   
9. RAD: index of accessibility to radial highways   
10. TAX: full-value property-tax rate per $10,000   
11. PTRATIO: pupil-teacher ratio by town   
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town   
13. LSTAT: % lower status of the population   
14. MEDV: Median value of owner-occupied homes in $1000's

Method:

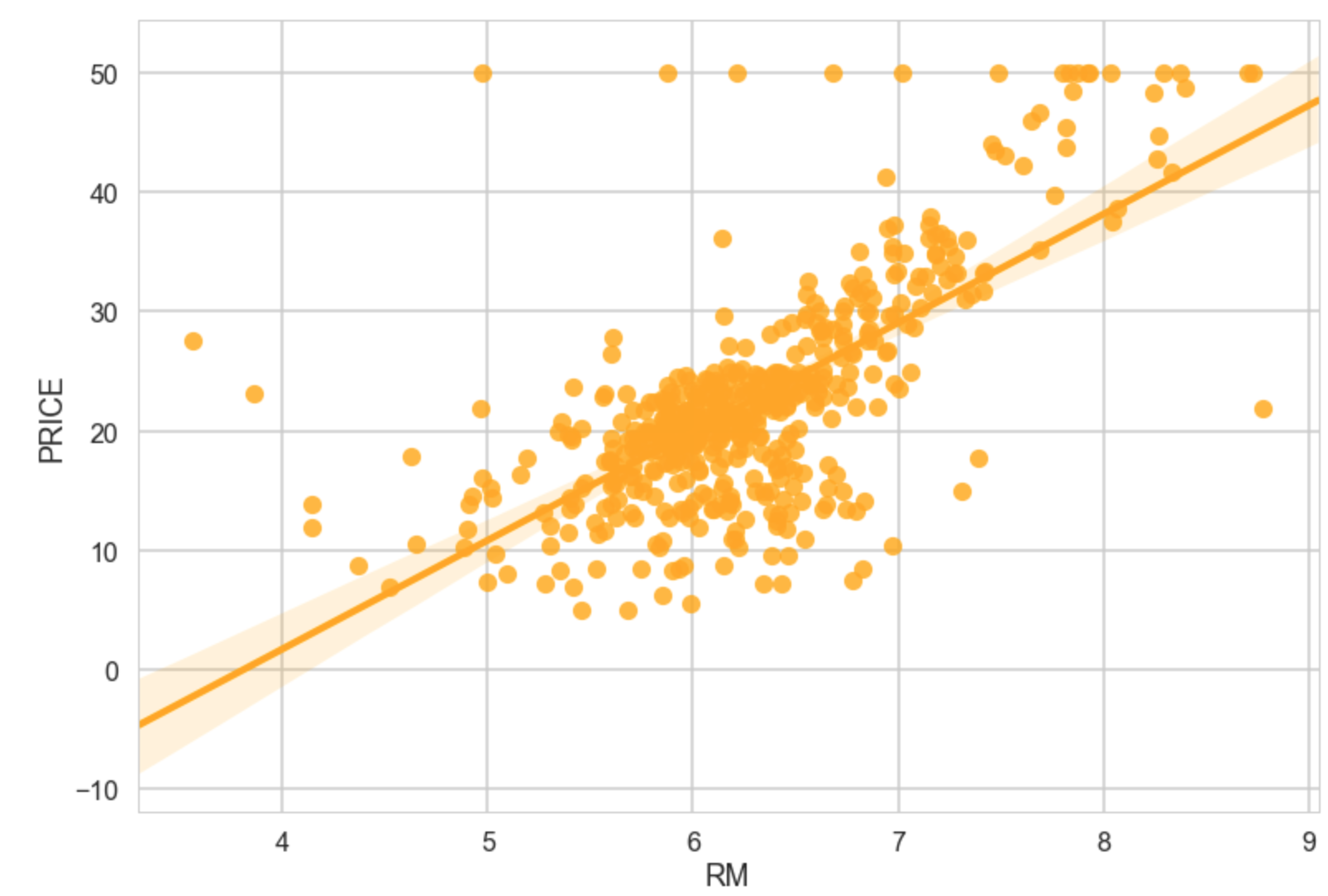
Checked all parameters and removed outliers from the data. Utilized sklearn library to develop linear modelling, checking the weightage of each variable on Price of houses provides clarity on the impacting factors. Further checking the MSE (Mean Squared Error) & R-squared re-enforces the correctness of model and fit.

Data Exploration:

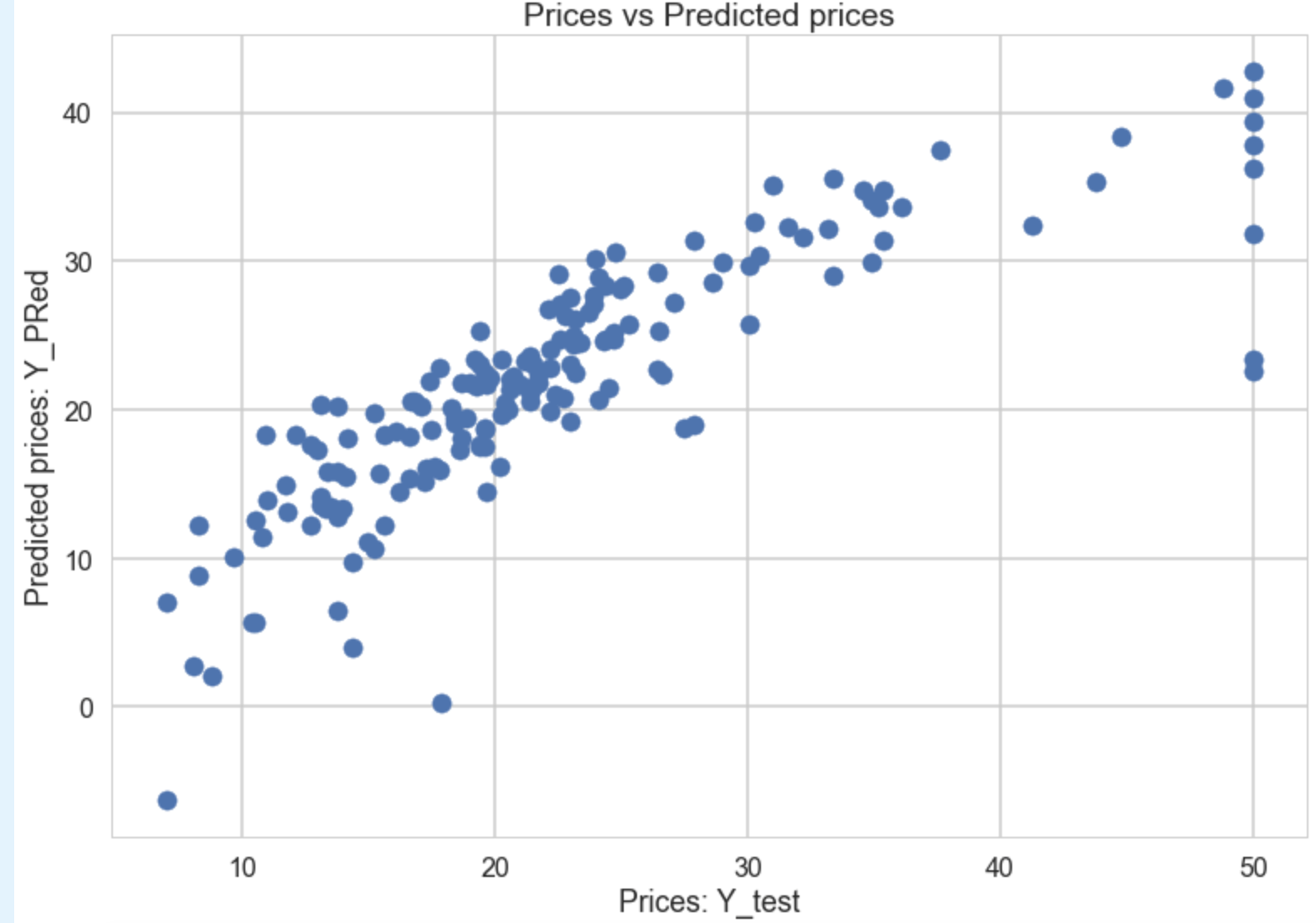
Plotting the Median of the target variable Price:



Plotting Price vs Number of Rooms in the House shows direct relation



Final Model Plotting Predicted House Prices vs Original Prices



Conclusion:

Below equation for linear modelling could serve as a base predicting the price of a house in Boston when all the other parameters are available.

**Y[PRICE] = 37.78 -1.13078590e-01*ZN +3.98065284e-02*INDUS -1.12702721e-01*CHAS +2.32300612e-01*NOX -1.12675573e+01*RM +2.88006535e+00*AGE -2.17309821e-02*DIS -1.25912181e+00*RAD +2.64204060e-01*TAX -1.28530861e-02*PTRATIO -8.29656073e-01*B +7.06606816e-03*LSTAT -3.80664041e-01MEDV**

Thus, we have an effective model in place, which can predict the price of a house in Boston.

Residual plot below re-confirms the Linear Modelling is the most appropriate solution for the data as the points are randomly displaced around the horizontal line.

