REPORT

SUMMARIZING THE DATA AND ITS PROCESSING

1. The Dataset and Its Features:

CRIM: Crime rate per capita by town.

ZN: Percentage of residential land zoned for lots over 25,000 sq.ft.

INDUS: Percentage 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: Percentage of owner-occupied units built before 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: Percentage of lower status of the population.

MEDV: Median value of owner-occupied homes in $1000's. (Output)

2. Data Preprocessing Steps:

Data Type Conversion: Changed data types from object to float for analysis and operations.

Handling Outliers: Identified and capped outliers in relevant columns.

Constant Column Check: Checked for and removed any constant columns.

Correlation Analysis: Analyzed and removed highly correlated columns to avoid multicollinearity.

Skewness Check: Assessed skewness in the data.

Data Transformation: Transformed columns to improve model performance.

Train-Test Split: Split the dataset into training and test sets.

Data Scaling: Applied Standard Scaler for scaling the data.

3. Model Training and Evaluation Results:

Linear Regression: Achieved an R^2 score of 81%.

XG Boost: Recorded an adjusted R^2 score of 82.08%.

Decision Tree Regressor: Recorded an adjusted R^2 score of 74.07%.

Gradient Boosting Regressor : Recorded an adjusted R^2 score of 84.11%.

Random Forest Regressor: Delivered an adjusted R^2 score of 84.17%.

Hyper-Parameter Tuning: Grid Search CV and Randomized Search CV on Random Forest resulted in an adjusted R^2 score of 82.08%, which was lower than the default Random Forest model. So it was not taken.

4. Interpretation of Model Performance and Coefficients:

- Concluded that the Random Forest model is the best, with an adjusted R^2 score of 84.17% .

- Identified important features using Random Forest and noted that some features contributed minimally to the output.

- After removing insignificant features, the adjusted R^2 score improved to 84.41%.

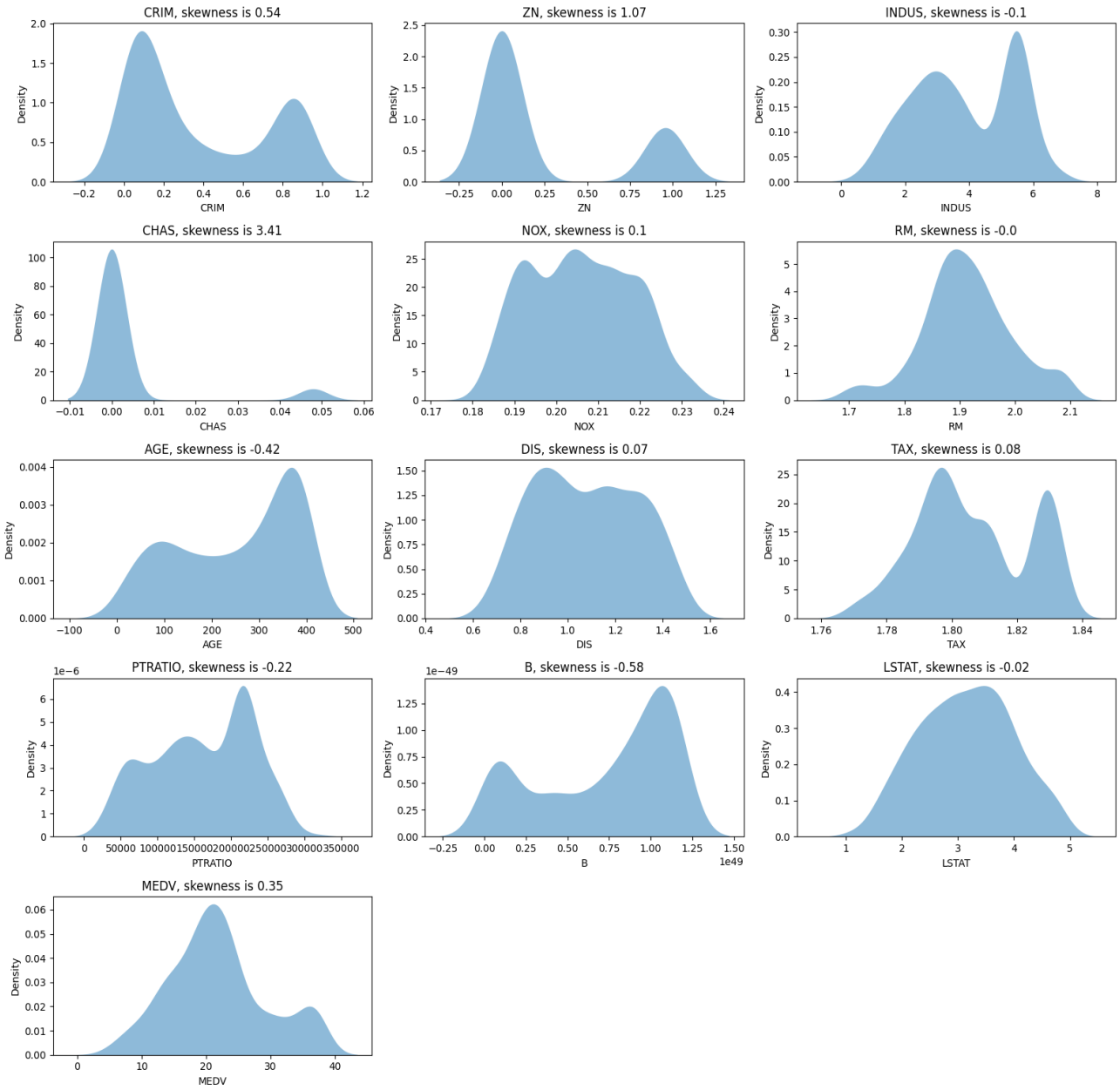
5. Challenges Faced:

The primary challenge was that hyper-parameter tuning degraded the model's performance.

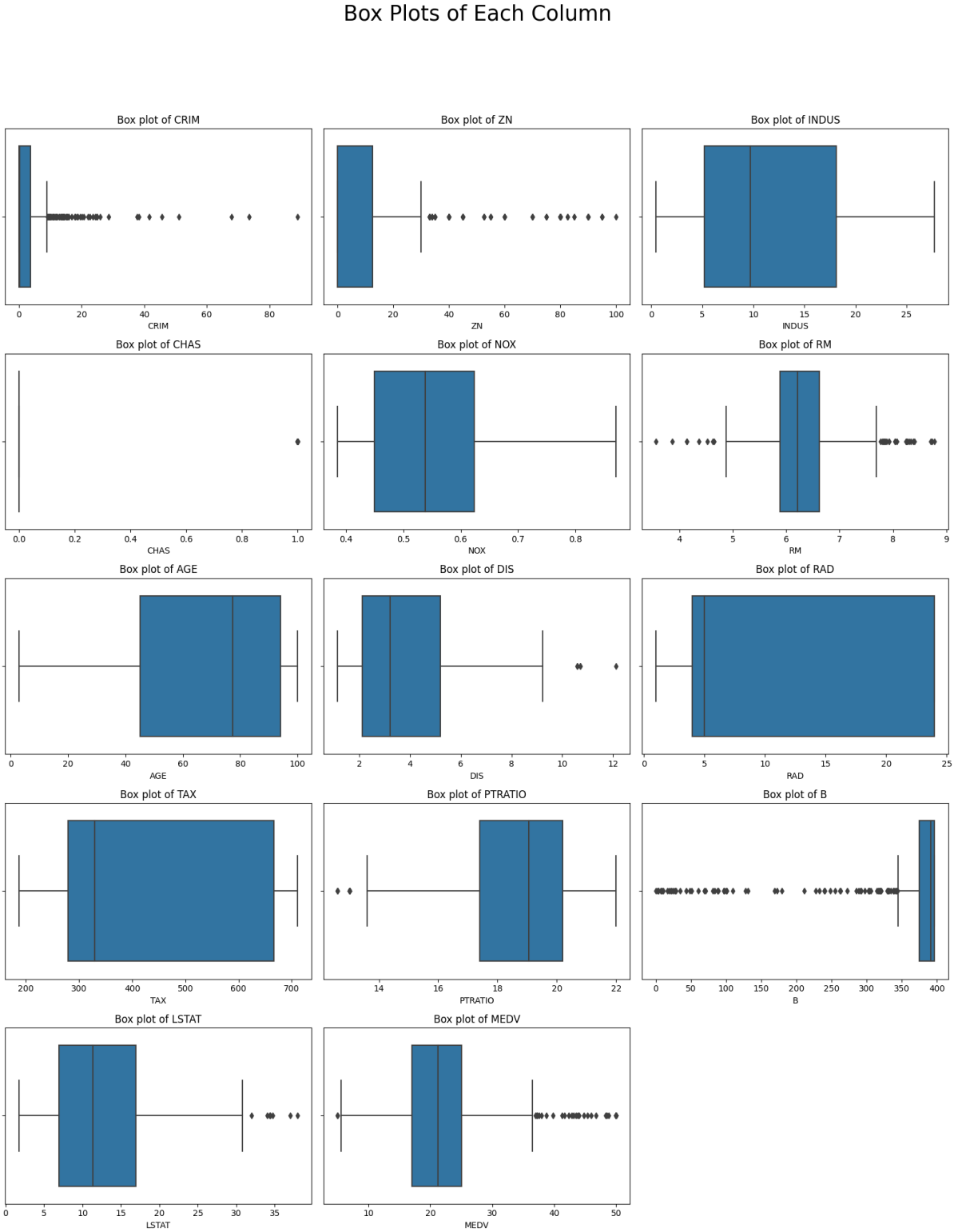
6. visualised results

Distribution before skewness removal

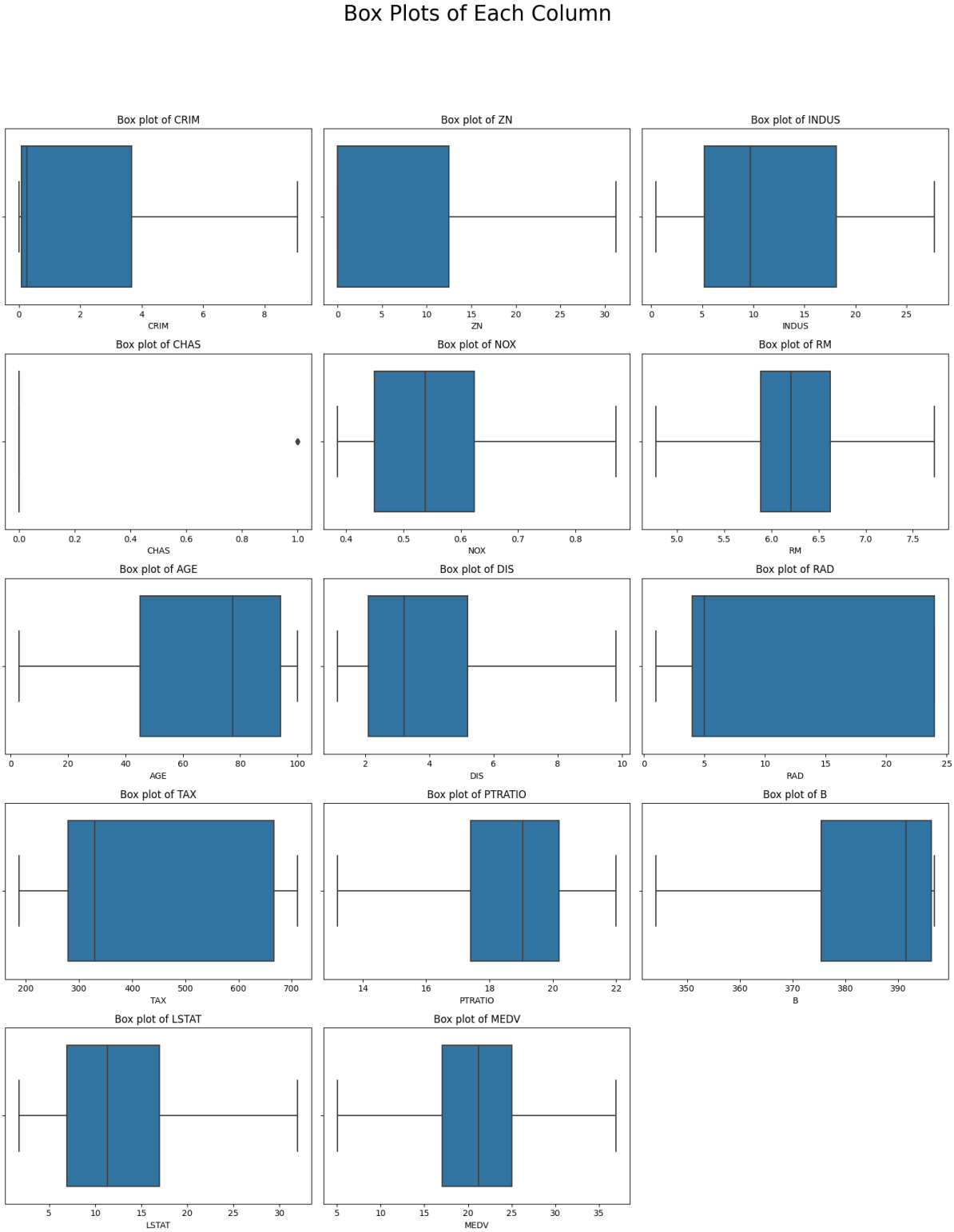
Distribution after skewness removal



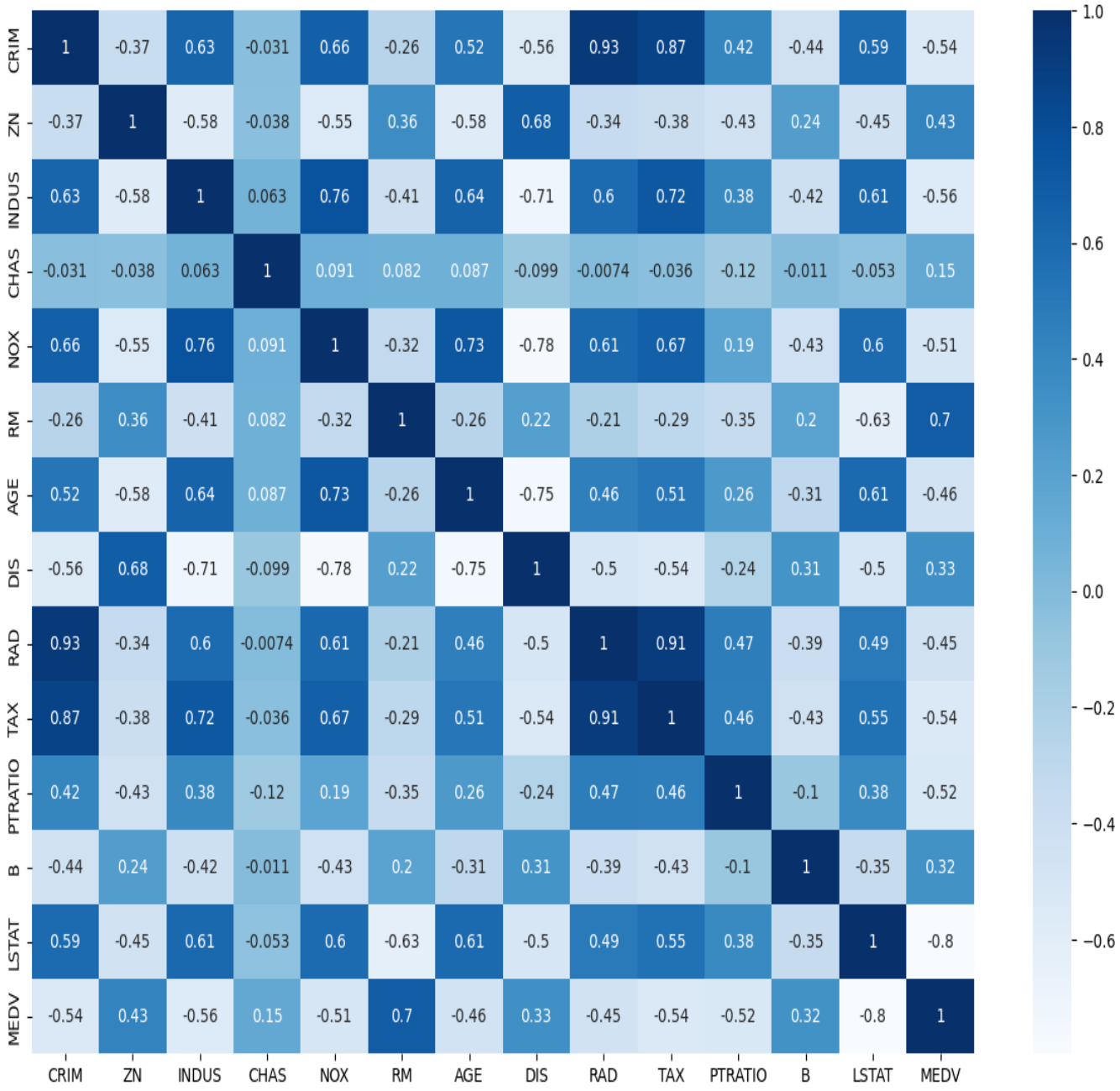
Boxplot before Outlier Removal



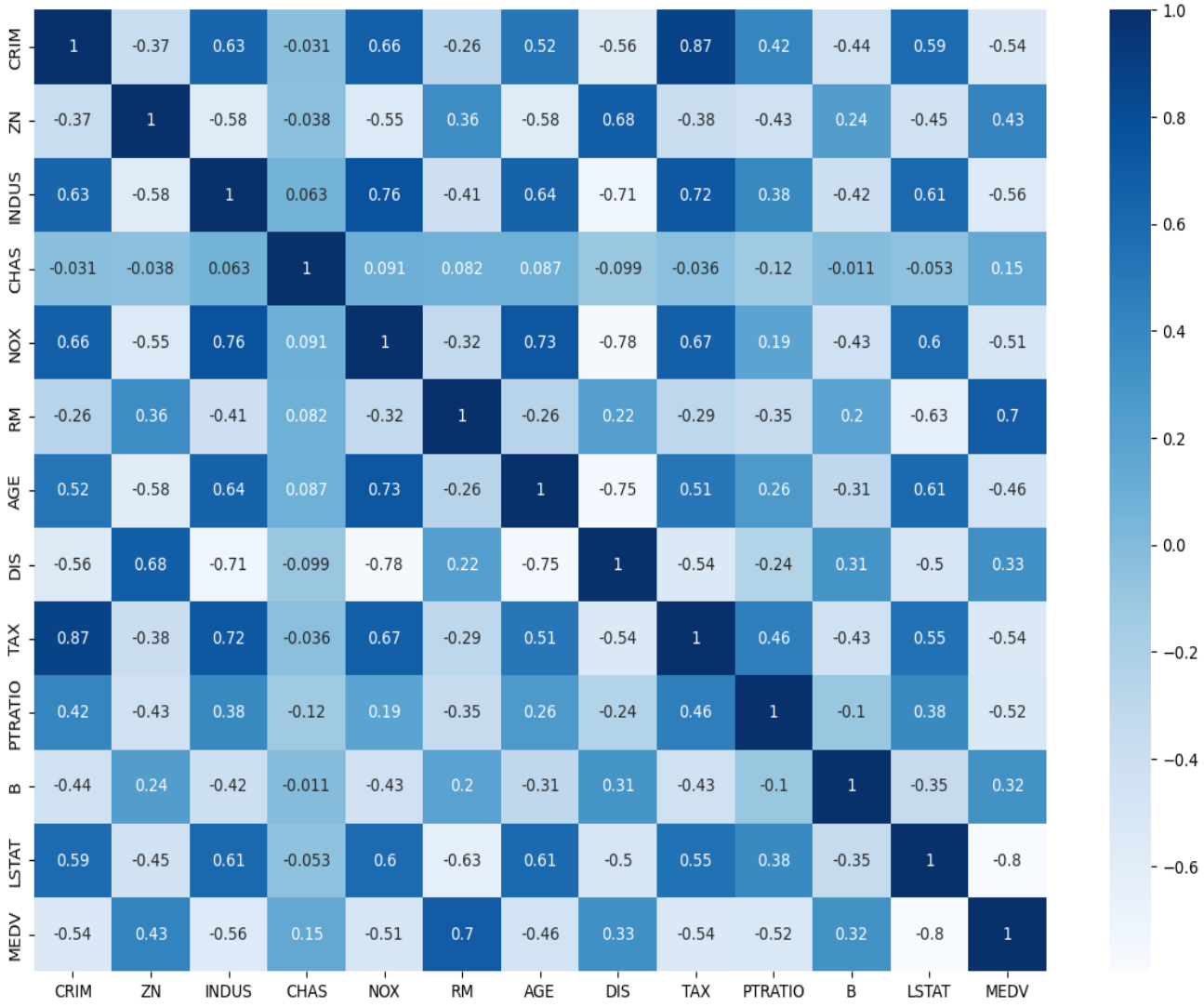
Boxplot after Outlier Removal



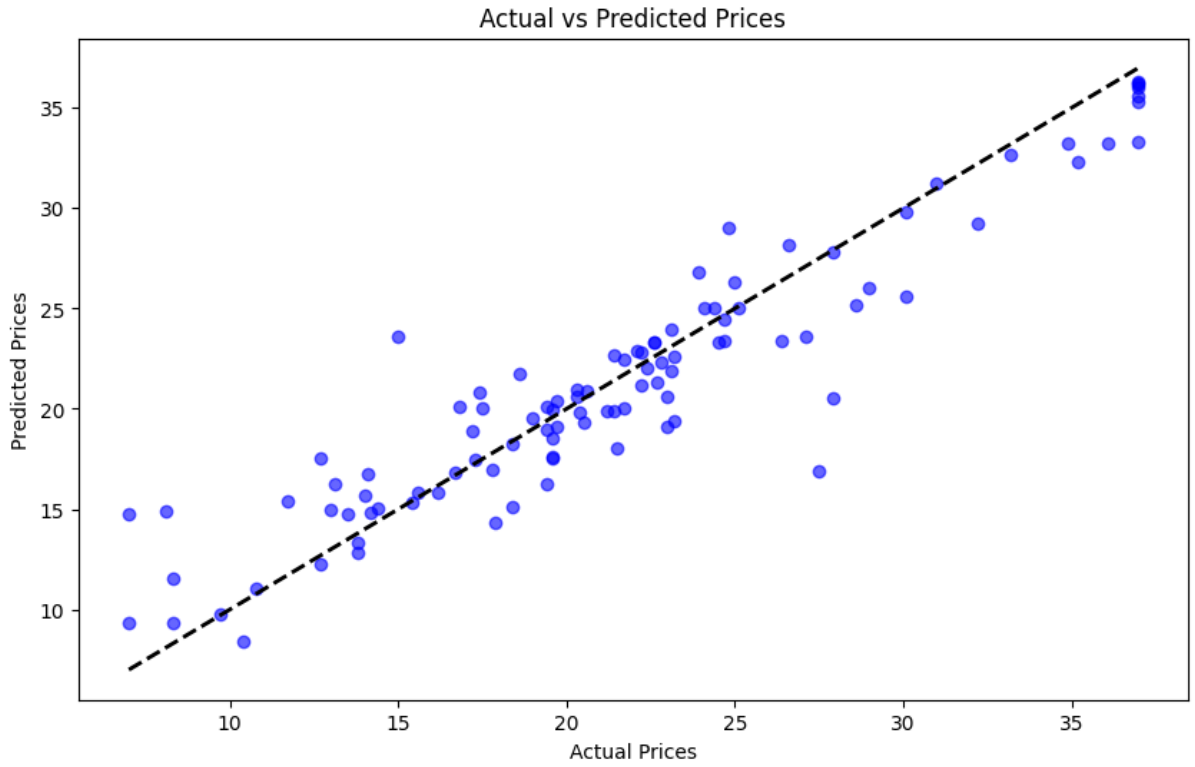
HeatMap before corelated feature removal



HeatMap after corelated feature removal



Actual vs Predicted price



NAME:- Pranav Parasar

SIC:- 21BCED77

Email: pranavparasar99@gmail.com