Predicting House Prices Using the Boston Housing Dataset

By : Abdul Wahab Aziz

Objective

The objective of this task was to explore the Boston Housing Dataset and build regression models to predict house prices. This involved:

1. Data preprocessing (normalization and splitting).
2. Visualizing the dataset to gain insights.
3. Implementing regression models (Linear Regression, Random Forest, and XGBoost) to compare performance.
4. Evaluating feature importance for tree-based models.

Dataset Overview

The Boston Housing Dataset contains the following features:

- **CRIM**: Per capita crime rate by town.
- **ZN**: Proportion of residential land zoned for large lots.
- **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 before 1940.
- **DIS**: Weighted distances to employment centers.
- **RAD**: Accessibility to radial highways.
- **TAX**: Property tax rate per $10,000.
- **PTRATIO**: Pupil-teacher ratio by town.
- **B**: 1000(Bk−0.63)21000(Bk - 0.63)^21000(Bk−0.63)2, where BkBkBk is the proportion of Black residents by town.
- **LSTAT**: Percentage of lower-status residents.
- **PRICE**: Median value of owner-occupied homes in $1000s (target variable).

Data Exploration

1. **Feature Distributions**:
   - Most features were not normally distributed. For example:

        - CRIM showed a right-skewed distribution, indicating higher crime rates in some towns.
        - RM (average number of rooms) exhibited a nearly normal distribution.
2. **Correlation Analysis**:
    - Strong positive correlations:
        - RM (+0.70+0.70+0.70) correlated positively with PRICE, indicating that homes with more rooms are typically more expensive.
    - Strong negative correlations:
        - LSTAT (−0.74-0.74−0.74) had a strong negative correlation with PRICE, showing that lower socioeconomic status is associated with lower home prices.
3. **Scatterplots**:
    - RM vs. PRICE: A clear upward trend, confirming that larger homes have higher prices.
    - LSTAT vs. PRICE: A downward trend, showing that areas with higher lower-status residents have lower prices.
4. **Outliers**:
    - Detected in features like CRIM, TAX, and DIS, which could affect model performance.

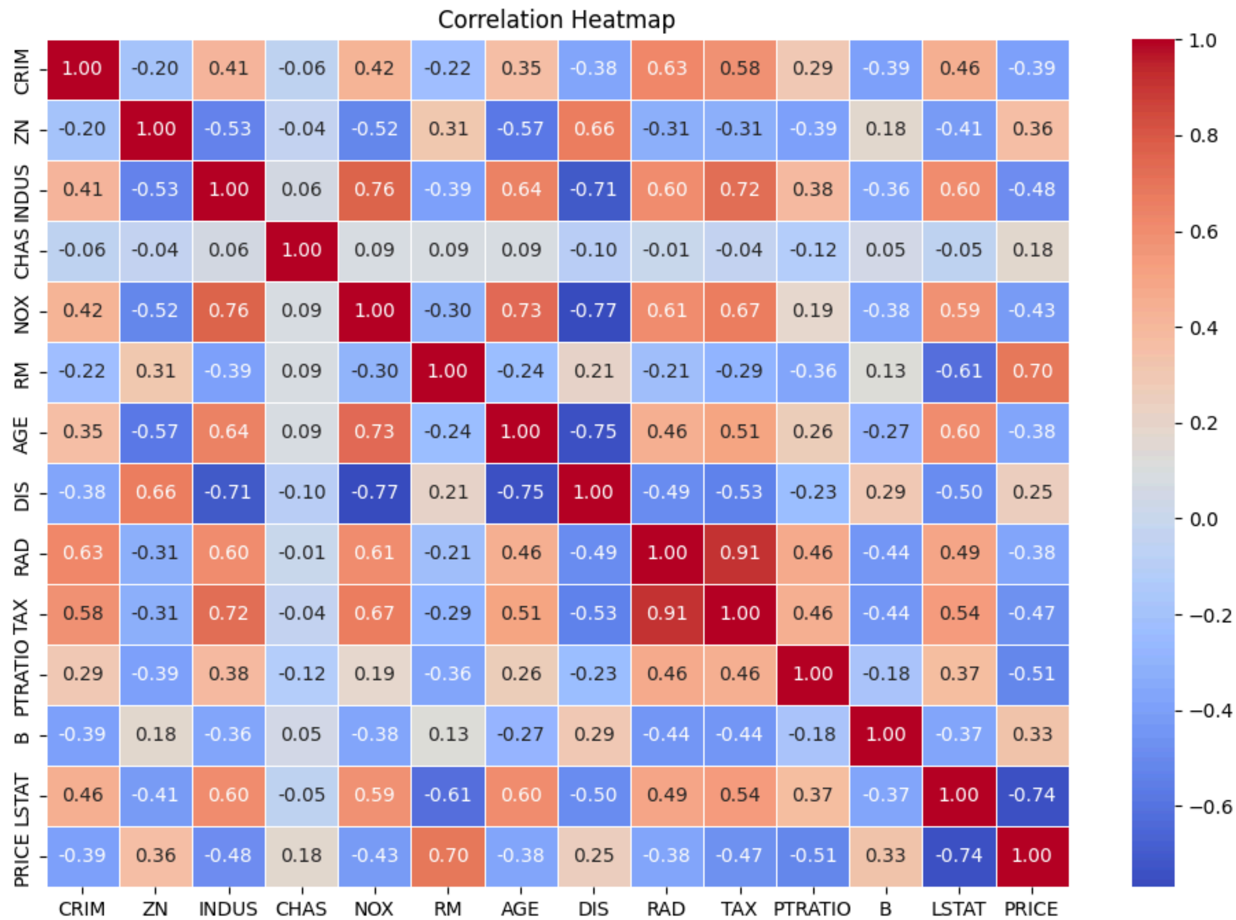
Insights

- **Key Drivers of House Prices**:
    - RM (number of rooms) and LSTAT (lower-status residents) were the most significant predictors.
    - DIS (distance to employment centers) also played a role, with higher distances generally leading to lower prices.
- **Challenges**:
    - Skewed distributions in features like CRIM and TAX might require transformations for better model performance.
    - Outliers could distort regression models and may need to be addressed.

Visualizations

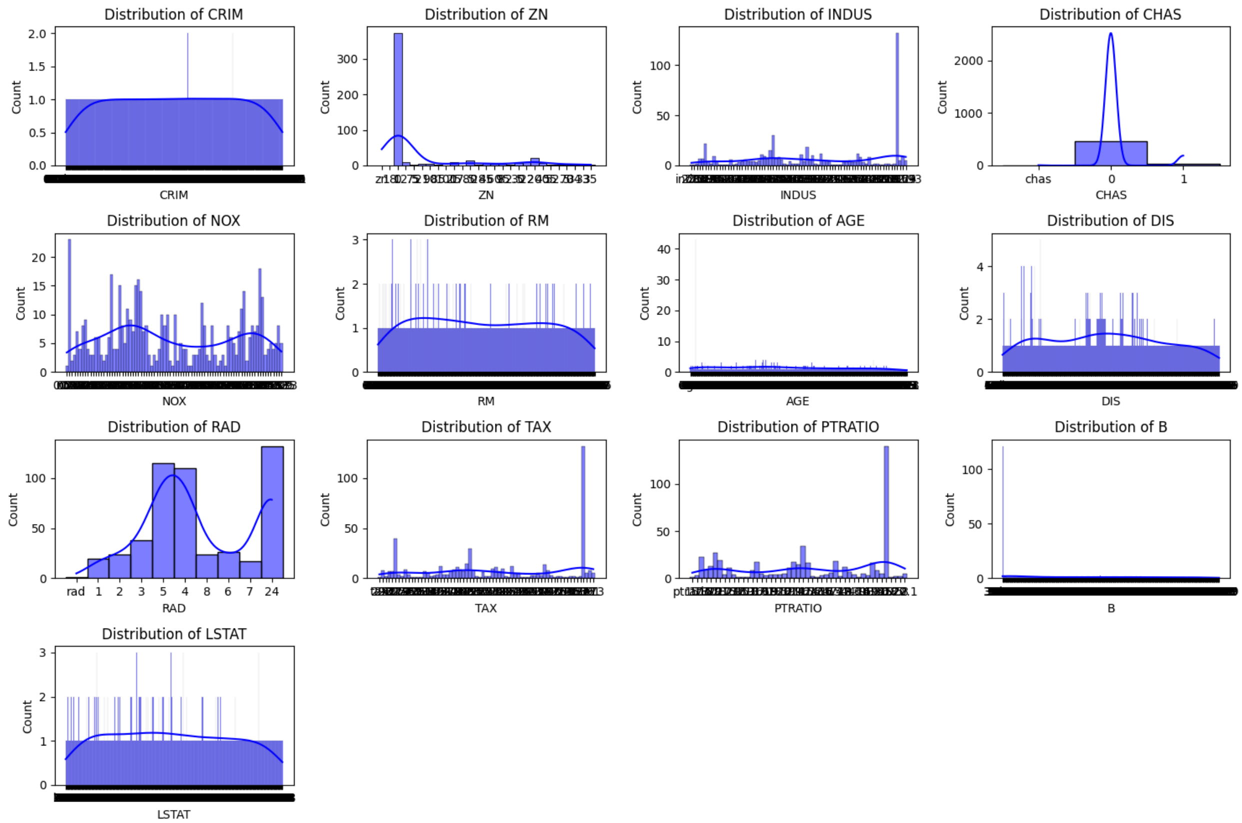
Correlation Heatmap: The heatmap highlighted:

- Positive relationships: RM, ZN.
- Negative relationships: LSTAT, PTRATIO.



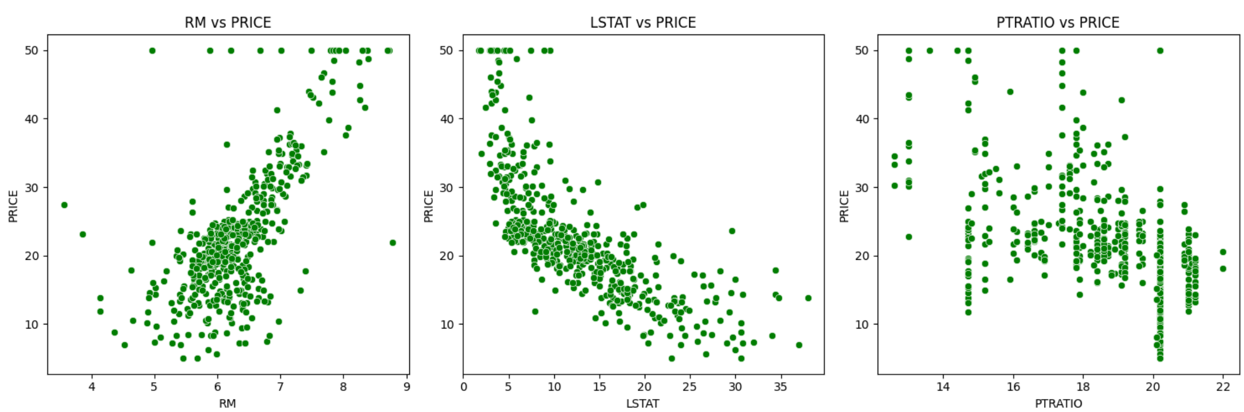
Feature Distributions:

- Provided insights into the spread and skewness of numerical features.



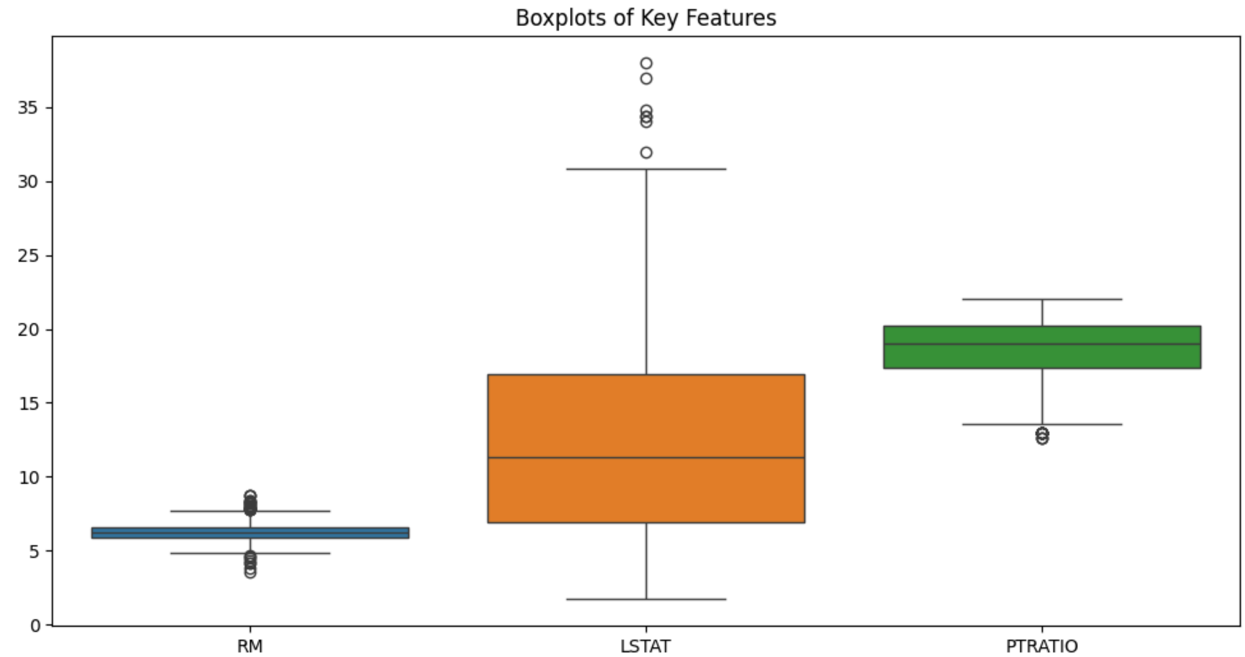
Scatterplots:

- Showed trends between features and PRICE.



Boxplots:

- Helped detect outliers in key features.



Conclusion

This exploration provided valuable insights into the dataset, such as:

- The strong relationships between features and house prices.
- The need to preprocess features with outliers or skewed distributions. These findings will inform model selection and feature engineering in subsequent steps.