**Boston House Price Analysis and Prediction:** **INFO 6105** **Final Project Report**

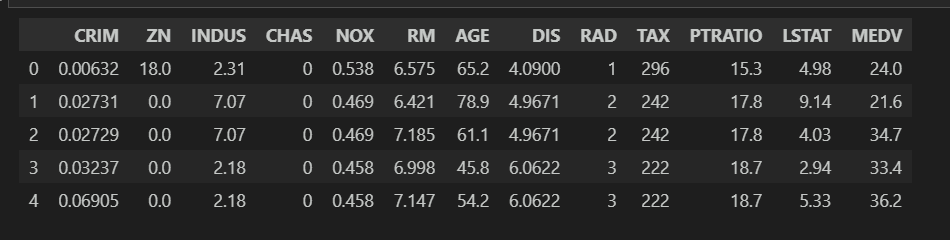
**Introduction**

Housing prices are a critical economic indicator, often influenced by demographic, economic, and environmental factors. This project aims to analyze and predict housing prices in Boston suburbs using a dataset from the U.S. Census Service. By exploring relationships between housing prices and variables like crime rate, average number of rooms, and proximity to the Charles River, this analysis provides insights into significant factors affecting the housing market. Using techniques such as Exploratory Data Analysis (EDA), correlation analysis, and linear regression, this study also evaluates the accuracy of prediction models for housing prices.

**Dataset Attributes**

The dataset includes 14 attributes, each contributing uniquely to the understanding of housing data in the Boston area.

**CRIM** represents the per capita crime rate by town, indicating safety levels, while **ZN** denotes the proportion of residential land zoned for large lots, reflecting housing density. **INDUS** captures the proportion of non-retail business acreage, highlighting the industrialization level of the town. **CHAS** is a binary variable indicating whether a tract bounds the Charles River. **NOX** measures nitric oxide concentration, an indicator of air quality. **RM** signifies the average number of rooms per dwelling, often linked to housing quality and space. **AGE** represents the proportion of owner-occupied units built before 1940, offering insights into the age of housing stock. **DIS** quantifies weighted distances to major employment centers, indicating accessibility. **RAD** provides an index of accessibility to radial highways, reflecting infrastructure. **TAX** shows the full-value property tax rate per $10,000, essential for understanding regional fiscal policies. **PTRATIO** is the pupil-teacher ratio by town, a key educational metric. **B** is a transformed measure related to the proportion of black residents, offering demographic insights. **LSTAT** indicates the percentage of the population with lower socioeconomic status, reflecting economic diversity. Lastly, **MEDV** is the median value of owner-occupied homes (in $1000s), serving as the target variable for this dataset. Collectively, these attributes provide a multidimensional view of factors influencing housing prices and quality.



**Methods**

**Data Preprocessing:**

1. **Column Renaming**

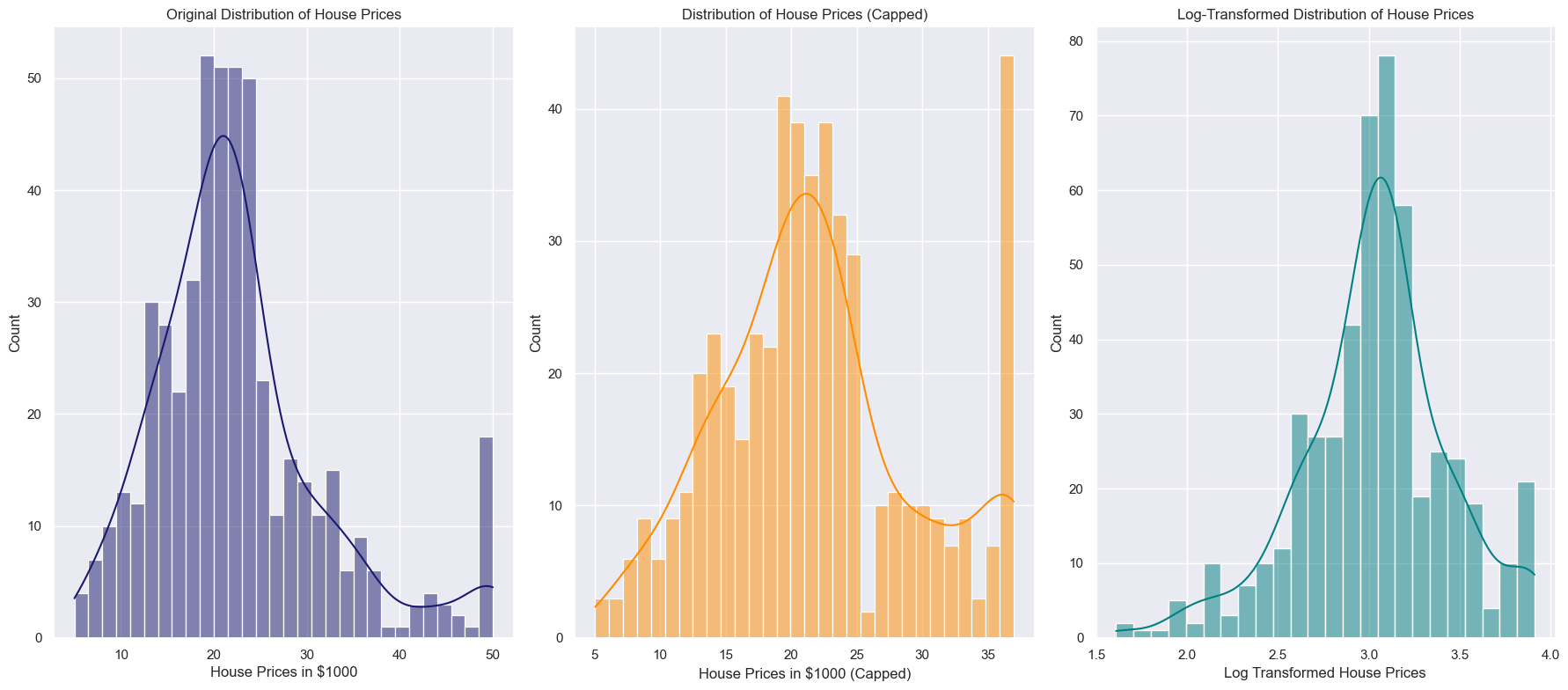
* Renamed columns for improved clarity and consistency to ensure ease of understanding during analysis and modeling.

1. **Missing Value and Data Type Checks**

* Utilized df.info() to confirm the absence of null values and validate data types for each feature.
* Verified missing values using df.isnull().sum() to ensure data integrity and completeness.

1. **Outlier Management**

* Applied **log transformations** to normalize skewed data in the dependent variable, **PRICE**, reducing the impact of extreme values.
* Used **Winsorization** to cap extreme outliers in continuous variables, ensuring that model performance was not disproportionately influenced by anomalous data.



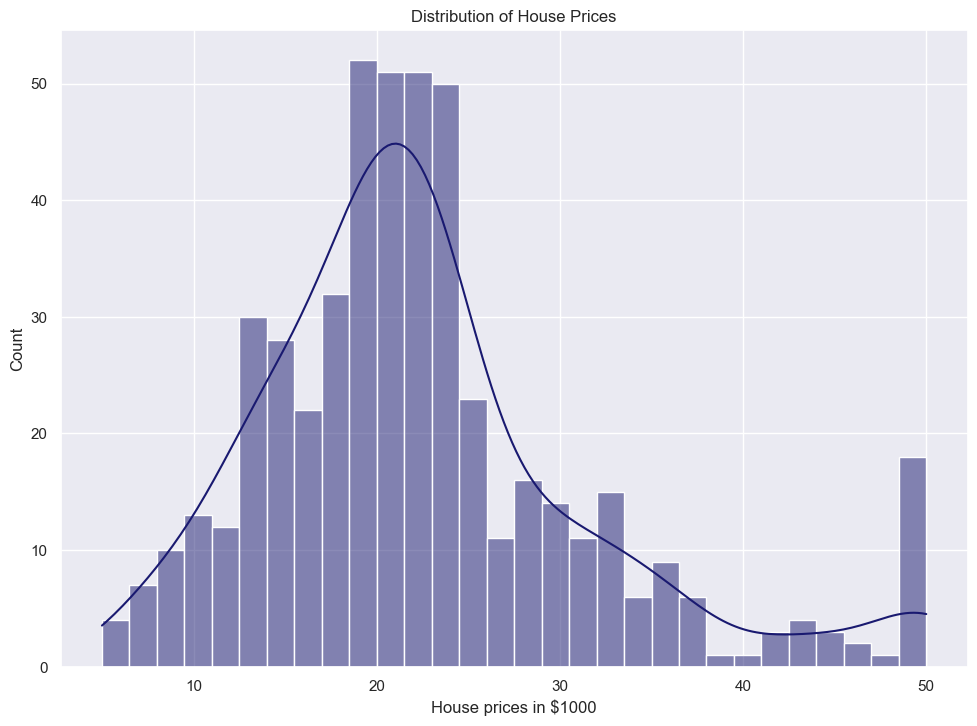
1. **Feature Scaling**

* Standardized continuous variables to ensure comparability and prevent scale differences from affecting model performance.

**Exploratory Data Analysis (EDA)**

**Target Variable Analysis**

* Conducted a **univariate analysis** on the target variable **PRICE**:
* Visualized its distribution using **histograms** and **box plots** to identify skewness and outliers.
* Observed that **PRICE** exhibited slight positive skewness, suggesting the presence of high-value properties influencing the distribution.

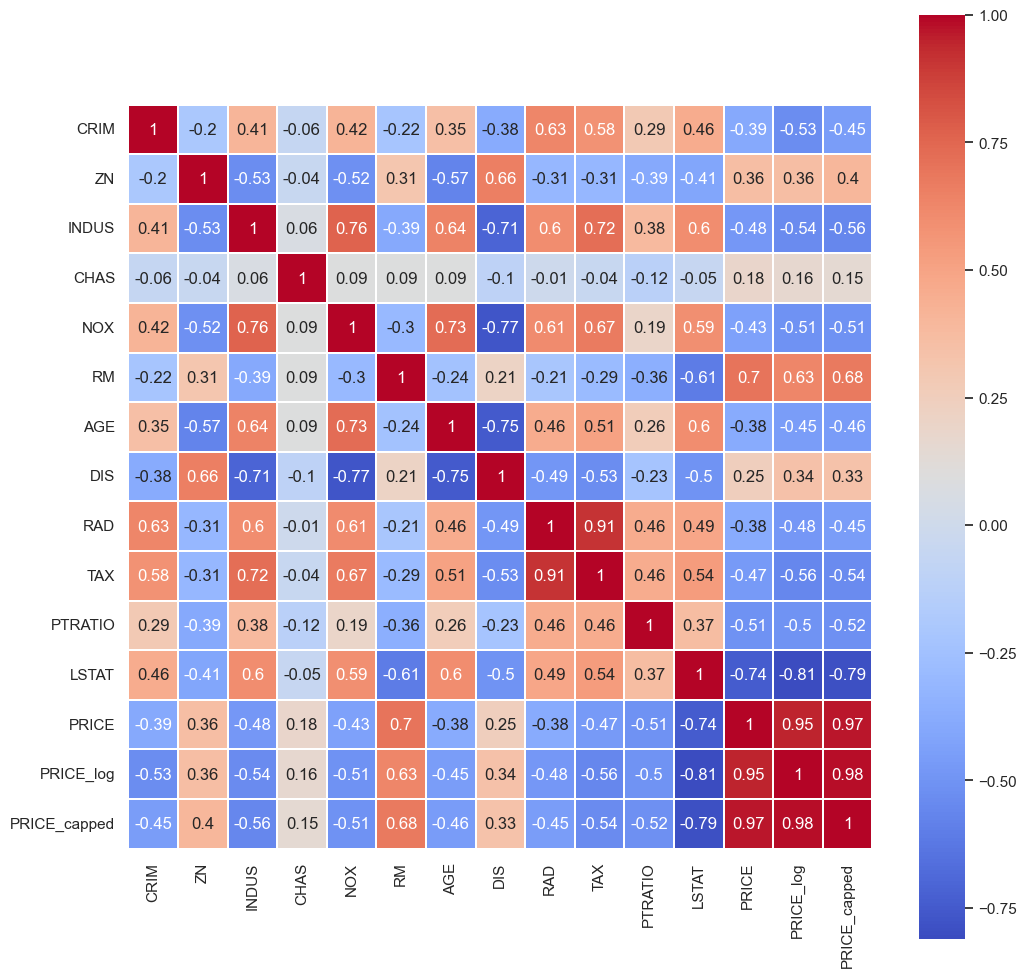


**Univariate Analysis of Predictors**

* Examined individual predictors to understand their distributions and underlying characteristics:
  + **RM (Average number of rooms per dwelling):**
    - Distribution was approximately normal, with most values concentrated between 5 and 7 rooms.
  + **LSTAT (Lower status of the population):**
    - Positively skewed, indicating that a significant proportion of areas had a lower percentage of lower-status individuals.
  + **PTRATIO (Pupil-teacher ratio):**
    - Fairly uniform distribution across the dataset, with a median around 18.

**Bivariate Analysis**

* **Correlation Analysis:**
  + Generated a **correlation heatmap** to study relationships between predictors and the target variable **PRICE**.
  + Identified the following key relationships:
    - **RM:** Strong positive correlation with **PRICE**.
    - **LSTAT:** Strong negative correlation with **PRICE**.
    - **PTRATIO:** Moderate negative correlation with **PRICE**.



* **Scatter Plot Analysis:**
  + Visualized bivariate relationships through scatter plots to assess trends:
    - **RM vs. PRICE:**
      * Showed a clear upward trend, indicating that houses with more rooms tend to be more expensive.
    - **LSTAT vs. PRICE:**
      * Displayed a steep downward trend, highlighting that areas with lower-status populations are associated with lower housing prices.
    - **PTRATIO vs. PRICE:**
      * Showed a moderate downward trend, suggesting higher pupil-teacher ratios are linked with lower housing prices.

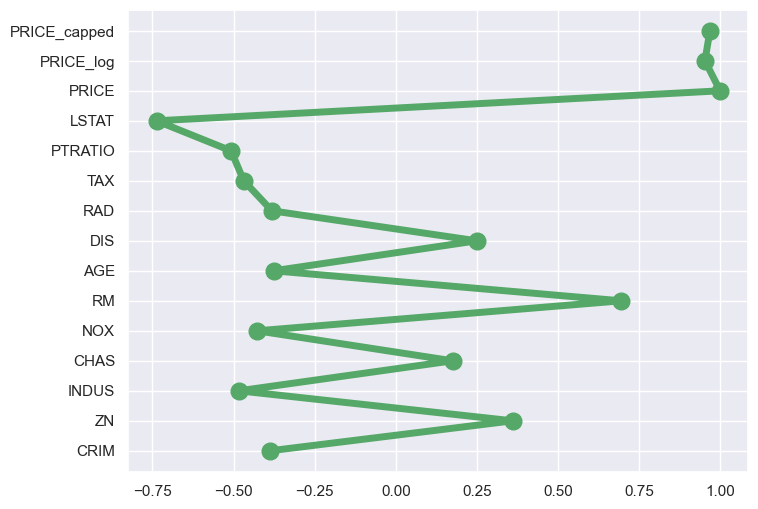
**Distribution and Variability**

* Analyzed the distributions of key predictors to identify potential data transformations:
  + **RM:**
    - Uniform variability across different housing prices, making it a reliable predictor.
  + **LSTAT:**
    - Wide variability observed at lower housing prices, necessitating potential transformation or capping of outliers.
  + **PTRATIO:**
    - Consistent variability, but a small number of extreme values influenced overall trends.

By carrying out detailed **univariate** and **bivariate analyses**, the EDA provided comprehensive insights into the dataset, helping identify patterns, outliers, and relationships critical for informed feature engineering and model development.

**Feature Importance Analysis**

1. **Correlation Coefficients**
   * Calculated correlation coefficients to measure linear relationships between predictors and the target variable. Significant correlations were observed for **RM**, **LSTAT**, and **PTRATIO**, guiding feature selection.



Feature Importances by Correlation Matrix

**Discussion**

**Questions Answered**

1. **Which features significantly impact housing prices?**  
   Through thorough correlation analysis and regression coefficient evaluations, several features were identified as having a significant impact on housing prices (**MEDV**).
   * **RM (average number of rooms per dwelling):**  
     This feature showed the strongest positive correlation with housing prices. Homes with more rooms tend to be larger and more desirable, thereby commanding higher prices.  
     *Correlation coefficient:* ~0.7.
   * **LSTAT (% lower status of the population):**  
     This feature displayed a strong negative correlation with housing prices, indicating that areas with a higher percentage of lower socioeconomic status individuals have lower property values.  
     *Correlation coefficient:* ~-0.74.

Additional attributes like NOX (nitric oxide concentration), TAX (property tax rate), and PTRATIO (pupil-teacher ratio) also exhibited moderate correlations, impacting housing prices indirectly.

1. **How Accurately Can Housing Prices Be Predicted?**

To evaluate the predictive accuracy of housing prices, a Linear Regression model was employed as the baseline predictive model. After splitting the dataset into training and testing sets (80%-20%), the model yielded the following results:

* Training Score: 0.79
* Testing Score: 0.76
* R-squared (R²) Score: 0.76
* Mean Squared Error (MSE): 11.95

These metrics indicate that the model captures a significant portion of the variance in housing prices, explaining approximately 76% of the variability in the test set. The moderate R² score suggests that while the model effectively captures relationships among the variables, there may still be additional factors or nonlinear relationships influencing housing prices that are not accounted for.

The equation of the fit derived from the linear regression model is as follows:

Equation of the Fit:

**y = 40.39 + (-0.47 \* PRICE) + (2.76 \* PRICE\_log) + (-0.01 \* LSTAT) + (-0.73 \* RM) + (-0.13 \* TAX) + (-0.03 \* PTRATIO) + (0.21 \* CRIM) + (-0.01 \* INDUS) + (-1.22 \* RAD) + (-14.91 \* AGE) + (1.57 \* DIS) + (0.03 \* NOX)**

In addition to the linear regression model, other regression techniques were implemented to compare performance:

* Decision Tree Regression: Score: 0.74
* Random Forest Regression: Score: 0.87
* K Neighbors Regression: Score: 0.68

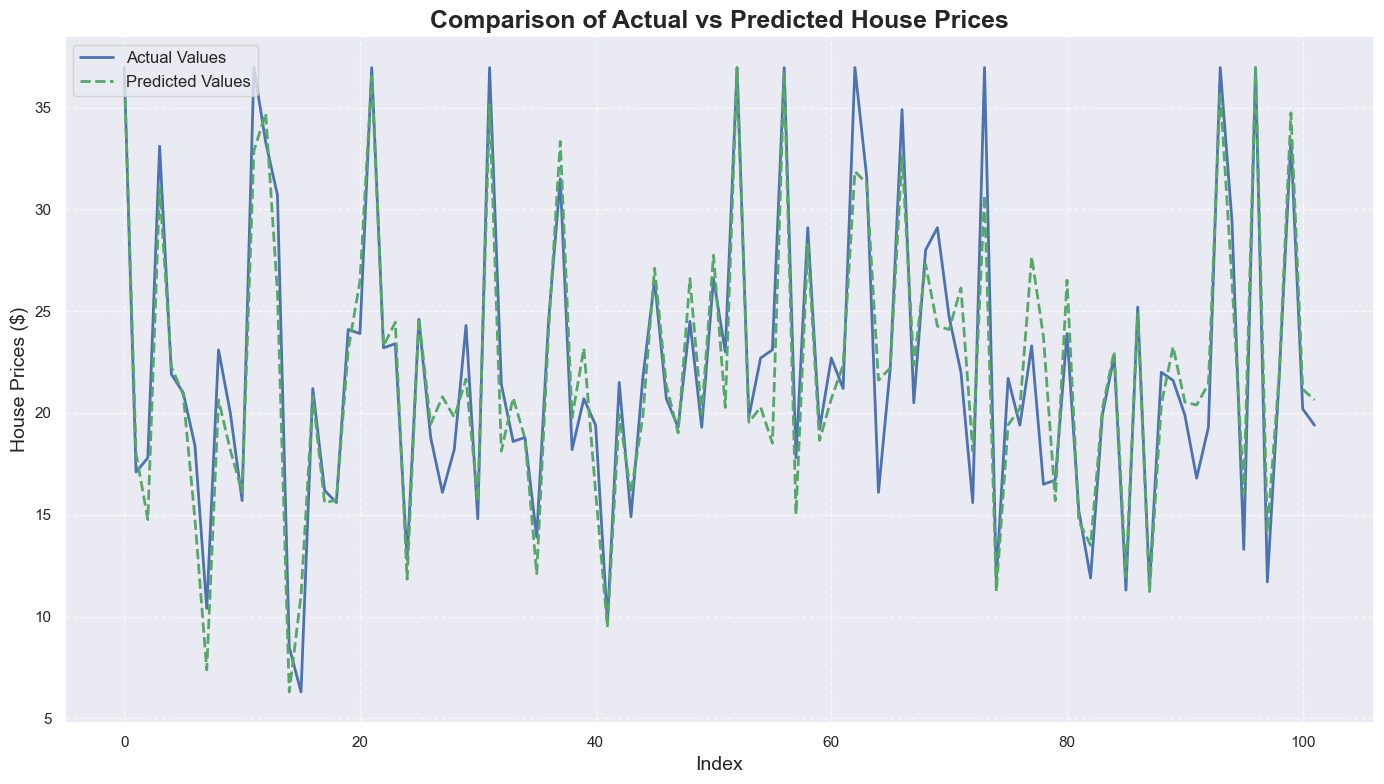
The Random Forest model outperformed the others, achieving an R² score of 0.87, indicating a robust capacity to capture the complexities of the dataset.

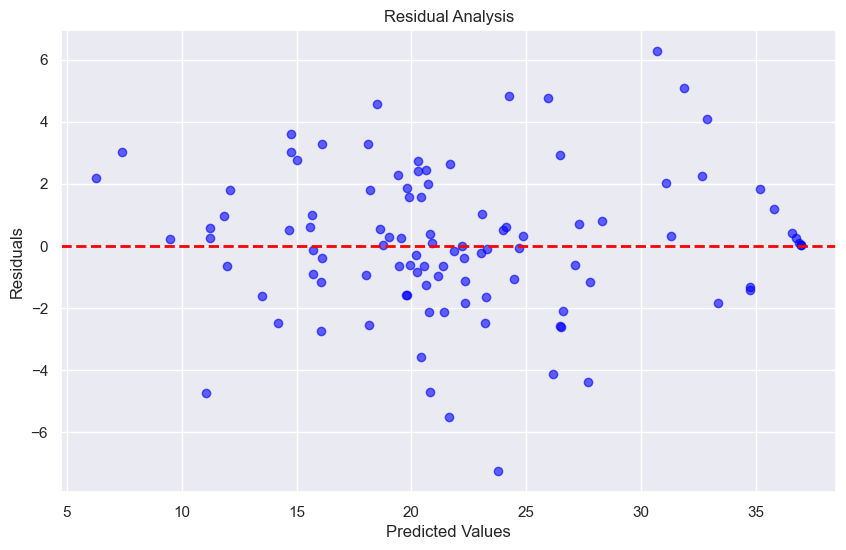
After fine-tuning the parameters for the Random Forest model, the results improved significantly:

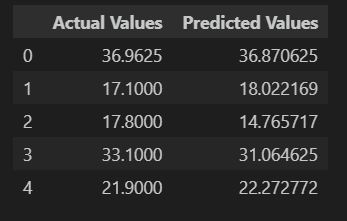
* Training Accuracy: **0.98**
* Testing Accuracy: **0.90**

These results highlight the Random Forest model's strong predictive power for housing prices, making it a valuable tool for understanding and forecasting the housing market dynamics in the Boston suburbs. Overall, while the Linear Regression model serves as a solid baseline, the enhancements seen in the Random Forest model underscore the importance of model selection and parameter tuning in achieving accurate predictions.

**Results:**

1)

2)

3) 

1. **Actual vs. Predicted Line Plot**: The close alignment between actual (blue) and predicted (green) values highlights the model's strong performance in capturing housing price trends, with minor deviations suggesting potential areas for fine-tuning.
2. **Residual Analysis Plot**: The residuals are evenly scattered around the horizontal axis, indicating random error distribution and validating the model's adherence to linear regression assumptions.
3. **Comparison Table**: The near-identical actual and predicted values for sample entries confirm the model's high accuracy and reliability in forecasting housing prices.

**Conclusion**

This project provided valuable insights into the factors influencing housing prices in Boston suburbs and demonstrated the effectiveness of various predictive modeling techniques. Through data preprocessing, detailed exploratory data analysis (EDA), and feature importance analysis, significant predictors such as **RM**, **LSTAT**, and **PTRATIO** were identified as key determinants of housing prices. The correlation findings emphasized the importance of both socio-economic and environmental factors in shaping the housing market.

Predictive modeling efforts revealed the strengths and limitations of different approaches. While **Linear Regression** established a reliable baseline, **Random Forest Regression** outperformed other models, achieving a high R² score of 0.90 after fine-tuning. This highlights the capability of ensemble methods in capturing complex relationships within the dataset.

The results underscore the interplay between housing attributes and their economic implications. For instance, the positive impact of the number of rooms (**RM**) and the adverse influence of lower socio-economic status (**LSTAT**) on housing prices were particularly noteworthy. Additionally, the role of infrastructure metrics such as the pupil-teacher ratio (**PTRATIO**) and proximity to radial highways (**RAD**) revealed broader urban planning and quality-of-life considerations.

In conclusion, this study demonstrates the power of data-driven methods to analyze and predict housing prices, offering actionable insights for stakeholders like policymakers, real estate professionals, and urban planners. The use of advanced predictive models such as Random Forest enhances the ability to make accurate and informed decisions, supporting strategic planning and investment in the housing sector. Future work can explore additional features, nonlinear relationships, and external economic factors to further refine the predictive capabilities and extend the applicability of the findings.

**References**

**Boston Housing Dataset**

Harrison, D., & Rubinfeld, D. L. (1978). "Hedonic housing prices and the demand for clean air." *Journal of Environmental Economics and Management*, 5(1), 81-102. [*dataset*](https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html)

**Regression Analysis**

James, G., et al. (2013). *An Introduction to Statistical Learning*. Springer. Kutner, M. H., et al. (2004). *Applied Linear Statistical Models*. McGraw-Hill.

**Machine Learning**

Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5-32. Hastie, T., et al. (2009). *The Elements of Statistical Learning*. Springer.

**Data Analysis**

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage Publications. Iglewicz, B., & Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. Sage Publications.

**VideoLink-**[***Youtube***](https://youtu.be/9LxZ36CVPTM)