Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna
College of Computer Studies

# Predictive Model for House Prices Using Linear Regression

## Introduction

This report outlines the development of a predictive model for estimating house prices using linear regression techniques. The model leverages various factors such as location, size, number of bedrooms, and age of the property to predict future house prices. The following sections detail the steps taken for data preprocessing, model development, evaluation, challenges faced, and the applicability of the model in real-world scenarios.

## Data Preprocessing

> Data Visualization and Exploration:
We began with exploratory data analysis (EDA) to understand the relationships between housing prices and various attributes. Key visualizations included:

- Correlation Matrices: To identify strong correlations between features.
- Scatter Plots: To visualize the relationship between house size and price, which showed a positive association.
- Histograms: To examine the distribution of house prices and other numerical features.

> Handling Missing Data:
Missing values were identified and addressed using the following techniques:

- Imputation: For numerical features, missing values were imputed using the mean or median.
- Dropping Rows: Rows with missing categorical variables were removed to ensure data integrity.

> Normalization and Encoding:
To ensure that all features were on the same scale, we normalized numerical features using Min-Max scaling. Categorical variables were encoded using one-hot encoding to transform them into a numerical format suitable for regression analysis.

## Model Development

> Model Implementation:
A linear regression model was created using the Scikit-learn module in Python. The dataset was divided into training (70%) and testing (30%) sets to properly assess model performance.

> Choosing Features:
Recursive Feature Elimination (RFE) was employed to identify the most important predictors, which helped in reducing model complexity and enhancing interpretability.

## Model Evaluation

> Performance Metrics:
The model was evaluated using the following metrics:

- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.
- R-squared: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables. The model achieved an R-squared value of approximately 0.85, suggesting a strong fit to the data.
- Adjusted R-squared: Adjusts the R-squared value for the number of predictors in the model, providing a more accurate measure of model performance.

> Visualization of Predictions:
A scatter plot of predicted prices against actual prices was created. The plot showed that most predictions were closely aligned with actual values, demonstrating the model's accuracy.

> Interpretation of Coefficients:
The coefficients of the regression model were analyzed to understand the impact of each feature on house prices. For example, an increase in size by 100 sq. ft. was associated with an increase in price by approximately $15,000.

## Challenges Faced and Solutions

> Data Quality Issues:
Challenges with inconsistent data entry and outliers were addressed through thorough data cleaning and the application of outlier detection techniques.

> Overfitting Concerns:
Cross-validation techniques were employed during model training to mitigate overfitting, ensuring the model performed well on unseen data.

> Feature Multicollinearity:
Multicollinearity among features was detected using the Variance Inflation Factor (VIF). Highly correlated features were removed to improve model stability.

## Visualizations and Plots

> Scatter Plot of Actual vs. Predicted Prices:
This visualization effectively illustrates the model's predictive power, showing a tight clustering around the line of equality.

> Correlation Matrix:
The correlation matrix displayed the relationships between features, highlighting strong correlations, particularly between size and price.

> Histogram of Residuals:
The histogram of residuals assessed the normality of residuals, confirming that they were approximately normally distributed, which is a key assumption of regression analysis.

## Conclusion

The developed linear regression model for predicting house prices demonstrates strong applicability in real-world scenarios, particularly for real estate agents and potential home buyers. It provides a reliable estimate based on key factors influencing house prices. However, the model has limitations, such as sensitivity to outliers and the assumption of linear relationships between features and the target variable. Additionally, the model's performance may vary with different datasets or in regions with distinct real estate dynamics. Future improvements could include exploring advanced regression techniques or incorporating additional features such as economic indicators or neighborhood characteristics to enhance predictive accuracy.