This project focuses on predicting house prices in Bengaluru using various machine learning models. The process can be summarized as follows:

### 1. Data Preprocessing

- \*\*Reading the Data\*\*: The dataset `Bengaluru\_House\_Data.csv` was loaded using Pandas.

- \*\*Initial Exploration\*\*: Inspected the first few rows and checked for null values and data types.

- \*\*Handling Missing Values\*\*: Filled missing values in categorical columns ('location', 'size') with the mode, and in numerical columns ('bath', 'balcony') with the median.

- \*\*Feature Selection and Engineering\*\*:

- Dropped columns that were deemed irrelevant based on domain knowledge and initial analysis ('area\_type', 'society', 'availability').

- Extracted the number of bedrooms ('bhk') from the 'size' column and then dropped 'size'.

- Cleaned the 'total\_sqft' column by handling ranges and non-numeric values.

### 2. Outlier Removal

- \*\*Price per Square Foot\*\*: Calculated a new feature 'price\_per\_sqft'.

- \*\*Removing Outliers\*\*: Removed data points with price per square foot values beyond one standard deviation from the mean for each location.

- \*\*BHK Outlier Removal\*\*: Further refined the dataset by removing properties with unusually low prices compared to properties with one less bedroom in the same location.

### 3. Data Transformation

- \*\*Feature Scaling and Encoding\*\*: Used `StandardScaler` for scaling numerical features and `OneHotEncoder` for encoding the 'location' feature.

### 4. Model Building and Evaluation

- \*\*Train-Test Split\*\*: Split the dataset into training and testing sets (80-20 split).

- \*\*Model Training and Evaluation\*\*:

- \*\*Linear Regression\*\*:

- R-square: 0.47

- MAPE: 0.85

- \*\*Random Forest Regressor\*\*:

- R-square: 0.72

- MAPE: 0.18

- \*\*AdaBoost Regressor\*\*:

- R-square: -0.25

- MAPE: 0.63

- \*\*Gradient Boosting Regressor\*\*:

- R-square: 0.71

- MAPE: 0.22

### 5. Model Selection and Saving

- \*\*Random Forest Regressor\*\* was selected as the best model based on performance metrics.

- \*\*Model Persistence\*\*: The trained Random Forest model pipeline was saved using `pickle` for future use.

### Summary

The project successfully cleansed and prepared the Bengaluru house price data, engineered relevant features, removed outliers, and built multiple machine learning models to predict house prices. The Random Forest Regressor was chosen as the final model due to its superior performance, and it was saved for deployment.