**INTRODUCTION**

**Real Estate Price Prediction Project**

This project is focused on building a machine-learning model to predict house prices in Bangalore, India. It involves several key steps, including data preprocessing, feature engineering, outlier removal, model selection, and Streamlet deployment. The project aims to provide a practical solution for homebuyers, sellers, and real estate enthusiasts in estimating property prices.

Dataset:- <https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data/code>

Technology and tools this project covers-

1. Python
2. NumPy and Pandas for data cleaning
3. Matplotlib for data visualization
4. Sklearn for model building
5. Jupyter notebook, visual studio code and PyCharm as IDE

Here's a breakdown of the major components and steps in this project:

1. **Data Collection**: The dataset used for this project is obtained from Kaggle. It includes various features related to real estate properties in Bangalore, such as location, total square feet area, number of bedrooms (BHK), number of bathrooms, price, and more.
2. **Data Preprocessing**:
   * Initial data exploration to understand the dataset's structure and characteristics.
   * Handling missing values by removing rows with missing data.
   * Creating a new feature, "bhk," by parsing the number of bedrooms from the "size" feature.
   * Cleaning the "total\_sqft" feature by converting it into a numerical format. This involves handling ranges and non-standard entries.
   * Creating a new feature, "price\_per\_sqft," to normalize prices based on square feet area.
   * Dimensionality reduction by categorizing less frequent location values as "other" to reduce the number of categories.
3. **Outlier Detection and Removal**:
   * Removing outliers based on business logic, such as properties with low square feet per bedroom.
   * Utilizing the mean and standard deviation to identify and remove price per square feet outliers for each location.
   * Further refining the dataset by removing properties where the number of bathrooms exceeds the number of bedrooms plus two.
4. **Model Building and Selection**:
   * The Linear Regression model is chosen as the primary predictive model for this project. The dataset is split into training and testing sets for model evaluation.
   * Cross-validation is applied to ensure model performance consistency.
   * A GridSearchCV process is used to find the best hyperparameters for the selected models. Other models like Lasso and Decision Tree Regression are considered and evaluated for their performance.

This project can be a valuable resource for individuals looking to explore real estate trends and estimate property prices in Bangalore, India. It combines data science techniques, machine learning, and interactive web development to create a user-friendly tool for real estate analysis.