

## Housing Price Predictor

### 1. Choice of Dataset

- a. The dataset we selected for this project will be sourced from Kaggle which contains extensive real estate data, including key features such as the number of bedrooms, bathrooms, square footage, and location. This dataset was chosen due to its structured nature and sufficient volume of data for meaningful model training.

### 2. Methodology

- a. Data Processes
  - i. Data cleaning
    1. Remove duplicate entries and irrelevant attributes.
    2. Handle missing values through imputation or removal, depending on the quality of data.
  - ii. Feature selection
    1. Keep the most relevant features such as: number of bedrooms, bathrooms, square footage, and location.
    2. Normalize numerical features to improve model stability.
    3. In later phases, encode categorical data such as different cities.
  - iii. Exploratory Data Analysis
    1. Identify and address outliers or skewed data
    2. Determine feature importance through statistical analysis
    3. Visualize the correlation between house prices and feature distributions
- b. Machine Learning Model
  - i. Linear regression
    1. Computationally efficient
    2. Assumes a linear relationship between features and house price
  - ii. Decision trees
    1. Captures nonlinear relationships
    2. More robust to outliers
- c. Model Training & Evaluation
  - i. The dataset will be split into three: 1) training, 2) validation, 3) testing.
  - ii. Moreover, cross-validation techniques will be used to assess generalizability.
  - iii. Evaluation metrics for regression models
    1. Mean Squared Error
    2. Mean Absolute Error

### **3. Application:**

Users will provide property details through a web form, which will include:

- i. Number of Bedrooms (numeric input)
- ii. Number of Bathrooms (numeric input)
- iii. Square Footage (numeric input)
- iv. Location
- v. Property Type
- i. Once the user submits their property details, the model will predict the estimated price.  
The output will be displayed as:
  - a. A numerical price estimate