**Problem Statement**:

The goal of this project is to analyze real estate data for houses in Bengaluru, with the goal of understanding pricing patterns by removing outliers to improve future analysis .

### **2. Tech Stack You're Using & Why You Selected It**

* **Python**: Because it has powerful libraries for handling data (like Pandas), performing calculations (NumPy), and visualizing results (Matplotlib).
* Python has a huge developer community. If I face any problems or challenges, there’s a vast amount of resources, documentation, and community forums available for help.
* Python offers all the tools I need for this project, from data cleaning to visualizing patterns in housing prices.
* **Jupyter Notebook** It allows interactive development, making it easy to see the results of each step of the data analysis process. You can quickly test and visualize your work.

### **3. To Give overall high level , how i'm going to solve the problem is:**

* **Load the Data**: Import the CSV file containing housing data.
* **Clean the Data:** Removing irrelevant columns, Handling missing values, Some locations have very few entries (less than 10), which can cause noise in the analysis. To manage this, we categorize these less frequent locations as 'rare'.
* **Feature Engineering**: Create new useful features price per square foot.
* **Remove Outliers**:Identify and remove unrealistic or extreme data points (e.g., properties with very high or low prices per square foot).

1)If each bedroom sqft<300 we removed it.

2) For any location properties with prices outside one standard deviation from the mean

are considered outliers and are removed.

3) If, for example, a 3 BHK is priced significantly lower than the average 2 BHK in the

same location, it is flagged as an outlier and removed.

. Calculate the mean price per square foot for each BHK in each location.

. Compare the price per square foot of higher BHKs with lower BHKs.

. If a higher BHK is priced significantly lower than expected, mark it as an outlier and

remove it.

* **Visualize the Data**: Use scatter plots and histograms to understand patterns and distributions.

We evaluated several models using grid search cv: Linear Regression, Lasso, and Decision Tree, based on their performance scores. The **Linear Regression** model gave the best score of **0.6997**, indicating it captured the relationship between features and price well, despite its simplicity.

**4. Demo:**

### **5.. Challenges Faced & How You Overcame Them**

**Challenge 1:**

Some entries in the dataset had ranges in the total\_sqft column (e.g., "2100-2850"). It was difficult to use such data directly.

**Solution**: I wrote a custom function to convert these ranges into the average value so they can be used in the analysis.

**Challenge 2:**

There were a lot of outliers (extreme values), such as houses with very low prices per square foot, which could skew the analysis.

**Solution**: I handled outliers by identifying properties that had unusual size or price and removed them from the dataset to ensure the analysis would be more accurate.

**Challenge 3:**

Locations with very few properties created noise in the analysis.

**Solution**: I grouped locations with fewer than 10 entries into a 'rare' category to reduce the noise and focus on more common locations.

### 

### 

### 