

## **Problem Statement (As Defined By The Client)**

### **Client Requirement**

- The Client, A Real Estate Firm, Wants To Develop A Pricing Model That Accurately Estimates The Market Value of Residential Properties Based on The Square Footage
- The Current Pricing Model (Predicted Price) Has Been Observed To Consistently Underpredict The Value of Properties When Compared To The Actual Market Prices, Especially For Larger Houses
- The Client Has Provided Historical Data of Residential Properties, Including Actual Market Prices And Square Footage Values, And Wishes To Identify The Shortcomings of The Existing Model.
- Client Seeks To Understand Why The Model Fails To Capture The Trend Observed in The Actual Prices And How A New Model Can Be Developed To Better Fit The Market

### **Specific Goals**

1. Identify And Explain The Trend Differences Between The Actual Property Prices And Predicted Prices Using Visual Representations
2. Evaluate The Accuracy of The Existing Linear Model For Predicting Property Prices And Identify The Reasons For High Bias
3. Propose A Revised Model That Fits The Market Data More Accurately

## **Problem Statement (As Defined By The Data Scientist)**

### **Title**

Evaluating And Improving The Predictive Accuracy of A Real Estate Pricing Model Using Machine Learning Techniques

### **Objective**

- To Analyze And Improve The Existing Pricing Model For Residential Properties That Currently Exhibits High Bias And Poor Fitting Characteristics
- The Project Aims To Identify Discrepancies Between Actual And Predicted Property Prices And Propose An Enhanced Model That Minimizes Prediction Error, Thus Improving The Alignment of Predicted Prices With Actual Market Values

### **Problem Description**

- The Current Predictive Model Used By The Real Estate Firm is Based on A Linear Formula  
**Predicted Price** = BasePrice + LinearCoefficient \* SquareFootage
- However, When Compared To The Actual Pricing Trend  
**Actual Price** = BasePrice + ScalingFactor \* (SquareFootage)<sup>2</sup>

### **Remarks on The Current Model**

- The Predicted Values Consistently Underestimate Property Prices, Particularly For Larger Square Footage Values
- The Quadratic Nature of The Actual Price Formula Indicates That Prices Increase Exponentially With Square Footage, While The Linear Model Only Captures A Direct Linear Relationship

- This Results in A Model With **High Bias**, Missing The Mark on Higher-Valued Properties And Failing To Accurately Represent Market Trends
- The Discrepancy Suggests That The Existing Linear Model is Insufficient To Capture The True Relationship Between Square Footage And Property Prices
- The High Bias Leads To A Consistent Underestimation, Indicating The Need For A More Complex, Non-Linear Model That Can Capture The Exponential Price Growth.

**Approach**

1. **Data Analysis**

Perform Exploratory Data Analysis (EDA) To Understand The Distribution of Square Footage And Corresponding Actual Prices

2. **Model Evaluation**

Plot And Analyze The Actual Vs. Predicted Prices To Visualize And Quantify The Bias

3. **Model Redesign**

Implement And Test Alternative Models (e.g., Polynomial Regression) To Capture The Non-Linear Relationship Between Square Footage And Prices

4. **Model Validation**

Evaluate The Revised Model’s Performance Using Appropriate Metrics Such As

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R-Squared Value.

**Expected Outcome**

- An Improved Pricing Model That Significantly Reduces Bias And Error, Accurately Predicting Property Prices Across Different Square Footage Ranges
- This Will Provide The Client With A More Reliable Tool For Estimating Property Values, Leading To Better Decision-Making in Property Transactions

**Sample Data**

Square Footage (sq_ft)	Actual Price (\$)	Predicted Price (\$)
1000	170,000.00	120,000.00
1500	320,000.00	45,000.00
2000	530,000.00	170,000.00
2500	800,000.00	195,000.00
3000	1,130,000.00	220,000.00
3500	1,520,000.00	245,000.00
4000	1,970,000.00	270,000.00
4500	2,480,000.00	295,000.00
5000	3,050,000.00	320,000.00

Variables Assumed	
Scaling Factor	0.12
Linear Coefficient	50.00
Base Price (Actual Price)	50,000.00