OPTIMIZING PROPERTY PRICING USING MATLAB MULTIPLE LINEAR REGRESSION

1. Overview

Client:

A regional real estate investment consultancy based in the United States

Objective:

To develop a reliable and interpretable pricing model for residential properties using MATLAB's regression capabilities. The goal was to identify undervalued or overpriced listings to support strategic purchase recommendations.

2. Background

The client routinely assessed property listings but relied on comparative heuristics rather than data-driven modeling. With an expanding portfolio and growing dataset of historical sales, they needed a predictive model that could guide buy/sell decisions more accurately across different states.

3. Data Summary

Dataset:

Residential sales data for 2,500 properties over 24 months

Key Variables:

Variable	Type	Description
Sale_Price	Continuous	Final sale price in USD (dependent variable)
Square_Footage	Continuous	Total built area in sq. ft.
Bedrooms	Integer	Number of bedrooms
Bathrooms	Integer	Number of bathrooms
Year_Built	Integer	Construction year
Zip_Code	Categorical	Location identifier (converted to dummies)
Garage	Binary	1 = Has garage, 0 = No garage

Has_Pool	Binary	1 = Pool included, 0 = No pool
Listing_Age	Continuous	Days listed before sale

4. Methodology

Software Used:

MATLAB R2023b with Statistics and Machine Learning Toolbox

Workflow:

1. **Data Preparation:**
    - Loaded and cleaned data using readtable() and rmmissing()
    - Converted categorical variables (Zip_Code) to dummy variables using dummyvar()
    - Standardized continuous predictors to compare coefficients
2. **Regression Model Setup:**
    - Built multiple linear regression using fitlm()
    - Model formula:
      Sale_Price ~ Square_Footage + Bedrooms + Bathrooms + Year_Built + Garage + Has_Pool + Zip_Code + Listing_Age
3. **Diagnostics and Validation:**
    - Checked multicollinearity using Variance Inflation Factor (VIF)
    - Evaluated model fit using R^2, RMSE, and residual plots
    - Conducted stepwise regression to test reduced models
    - Split dataset: 80% training, 20% testing
4. **Report and Interpretation:**
    - Exported results using exportgraphics() and writetable()
    - Created interpretation section explaining key coefficients and their real-world meaning

5. Key Results

Metric	Value
Adjusted R^2	0.896
RMSE on Test Data	$18,400
Most Impactful Predictor	Square_Footage
Significant Amenities	Garage, Pool
Overpricing Flagged (Top 5%)	42 properties

Insights:

- Properties with **pool and garage** sold on average **$27,000 more** than similar properties without
- Properties in Zip Codes 90001 and 90003 were consistently **undervalued by $15–20k** based on model predictions
- Listings older than 60 days had a lower average sale price by **6.2%**

6. Visual Outputs (MATLAB):

- Coefficient plot with confidence intervals
- Residual vs. fitted plot (checked for homoscedasticity)
- Actual vs. Predicted price scatter
- Geographic price prediction map (zip-level)

7. Deliverables

- Final regression model in .mat format with coefficients
- Custom MATLAB functions for data import, preprocessing, and prediction
- 17-page professional report including:
    - Model development summary
    - Full regression table with p-values and CIs
    - Business interpretations
    - List of over- and under-valued properties

- Executive dashboard preview:
    - Sorted listing suggestions
    - Price deviation alert system

8. Application & Outcome

- Client integrated the model into their investment dashboard
- Used prediction intervals to benchmark new listings for investment scoring
- Achieved 93% pricing accuracy on test properties
- Identified three below-market deals that were secured within 60 days post-deployment

9. Strategic Value Delivered

- Replaced subjective price benchmarking with **objective, data-driven model**
- Delivered **interpretable, customizable MATLAB script**
- Provided **strategic buy/sell guidance** based on pricing deviation
- Helped establish client reputation for accurate market valuation forecasting