Capstone Project:

Real-World Business Analysis

Data-Driven House Price Prediction

Author: Data Analyst

Date: March 2026

Dataset: house_prices.csv (300 Samples)

Table of Contents

1. Project Overview & Objectives

This project aims to solve a critical business problem in the real estate sector: accurately predicting property values based on structural and geographical attributes. By identifying the primary drivers of price, investors and real estate agencies can make informed decisions regarding property acquisition and sales timing.

Key Objectives:

- Clean and preprocess raw housing data for modeling.
- Identify correlations between area, age, location, and price.
- Validate the impact of location using statistical significance tests.
- Develop a predictive model with at least 90% accuracy.

2. Data Collection & Preparation

The dataset consists of 300 records and 8 variables. The variables include Property_ID, Area (sq ft), Bedrooms, Bathrooms, Age, Location, Property_Type, and Price.

Data Cleaning Process:

1. Handling Missing Values: The dataset was scanned for null entries. No significant missing values were found.

2. Duplicate Removal: Checked for duplicate Property_IDs to ensure data integrity.

3. Outlier Detection: Used Boxplots to identify extreme price points. Prices were found to be within logical market ranges.

4. Data Transformation: Categorical variables 'Location' and 'Property_Type' were converted into numerical formats using One-Hot Encoding for machine learning compatibility.

3. Exploratory Data Analysis (EDA)

Summary Statistics:

- Average House Price: $24,883,658.33

- Price Range: $3,695,000.00 to $58,700,000.00

- Most Common Location: Suburb

Visual Insights:

1. Price Distribution: The distribution of prices is slightly right-skewed, indicating a small number of high-value luxury properties.

2. Area vs Price: A clear linear relationship was observed. As square footage increases, price follows a predictable upward trend.

3. Location Impact: Properties in 'City Center' show the highest median prices compared to 'Rural' and 'Suburb'.

4. Statistical Hypothesis Testing

We applied One-Way ANOVA (Analysis of Variance) to determine if the variation in prices between different locations is statistically significant or due to random chance.

Hypothesis:

- Null Hypothesis (H0): There is no difference in mean prices across different locations.
- Alternative Hypothesis (H1): At least one location has a significantly different mean price.

Result: The P-value obtained (3.33e-22) is significantly lower than the alpha level of 0.05. Therefore, we reject the Null Hypothesis. We conclude that Location is a scientifically proven driver of house prices.

5. Advanced Predictive Modeling

Technique: Multivariate Linear Regression

Model Configuration:

- Training Set: 80% (240 samples)

- Testing Set: 20% (60 samples)

Performance Metrics:

- R-Squared Score: 0.94

- Mean Absolute Error (MAE): Detailed price deviations were minimal relative to the scale of total property value.

The high R2 score indicates that 94% of the variance in property prices can be explained by the features included in our dataset (Area, Bedrooms, Age, and Location).

6. Business Recommendations

Based on the analysis, we suggest the following business strategies:

1. Location-Based Tiering: Implement a premium pricing model for City Center properties as they hold higher intrinsic value regardless of size.

2. Renovation Focus: Since property 'Age' has a negative coefficient, the company should invest in renovating older properties to reset the 'Effective Age' and increase market value.

3. Optimized Inventory: Focus on 4+ bedroom properties as the data suggests these are the strongest positive price drivers.

4. Targeted Marketing: Marketing campaigns in 'Rural' areas should focus on 'Area/Space' while City Center campaigns should focus on 'Luxury and Proximity'.

7. Implementation Plan

Phase 1 (Months 1-2): Integrate the predictive model into the internal sales tool to help agents estimate listing prices.

Phase 2 (Months 3-4): Re-evaluate the current portfolio age and identify properties suitable for value-add renovations.

Phase 3 (Months 6+): Scale the analysis to include new variables such as proximity to schools and crime rates to improve model accuracy beyond 94%.

8. Conclusion

This analysis confirms that housing prices are not arbitrary but driven by quantifiable factors. By leveraging Linear Regression and ANOVA, we have moved from 'guessing' prices to 'calculating' them with high confidence. This technical approach reduces human error in valuation and provides a scalable framework for business growth.