PREDICTING HOUSING PRICES USING MULTIPLE LINEAR REGRESSION IN R

Client Overview:

A mid-sized real estate advisory firm in Boston, Massachusetts approached me to develop a predictive model for property pricing. Their goal was to use historical transaction data to build an internal tool that could guide sellers on reasonable asking prices while also identifying undervalued investment opportunities.

Problem Statement:

Despite having access to property listings and previous sale records, the firm lacked a consistent, data-driven method to price properties. Manual valuation led to inconsistent recommendations, pricing errors, and missed opportunities. They needed a model that could accurately predict property prices based on specific attributes.

Solution Approach:

I proposed a Multiple Linear Regression model using R to quantify the influence of key factors on property prices. The project was broken down into the following phases:

1. Data Collection and Preprocessing

- Imported 5 years of property transaction data (CSV format) covering Boston's downtown districts.
- Cleaned and preprocessed the data: handled missing values, removed multicollinear predictors, converted categorical variables into dummy variables, and standardized units (e.g., square footage).
- Final dataset included ~2,000 records with 12 variables.

2. Variable Selection

- Independent variables
  included: Square_Footage, Num_Bedrooms, Num_Bathrooms, Zip_Code (dummy coded), Year_Built, and Lot_Size.
- Target variable: Sale_Price.

3. Model Building

- Built and tested several regression models in R using the lm() function.
- Applied stepwise selection (AIC) to optimize the combination of predictors.
- Checked for multicollinearity using Variance Inflation Factor (VIF).
- Detected and handled influential points using Cook's Distance and leverage plots.

4. Model Diagnostics and Validation

- Verified assumptions:
    - **Linearity:** Residual vs Fitted plots
    - **Normality:** Histogram and Q-Q plot of residuals
    - **Homoscedasticity:** Breusch-Pagan test
- Model achieved an **Adjusted R^2 of 0.81** on training data and **RMSE of \$17,500** on test data.

5. Reporting and Insights

- Delivered a professional report (in R Markdown, exported to PDF and HTML) including:
    - Visualizations: scatterplots, regression lines, coefficient plots
    - Interpretation of coefficients: e.g., each additional 100 sq ft increased price by ~\$25,000 in downtown Boston
    - Summary table of model performance metrics
    - Recommendations on how to integrate the model into the firm's existing CRM tool

Business Impact:

- Enabled agents to quote data-backed asking prices within a confidence interval.
- Reduced pricing disputes with clients by providing transparent, quantitative justification.
- Helped identify 3 neighborhoods where predicted prices were consistently higher than list prices—triggering a re-evaluation of those markets.

Tools Used:

- **Software:** R (RStudio)
- **Libraries:** dplyr, ggplot2, car, caret, broom, MASS
- **Reporting:** R Markdown