An Analysis for Modeling in Housing Prices

Yuqi Shi

Introduction

This report investigates the determinants of housing prices using a dataset of property characteristics and market-level variables. The primary objective is to identify key drivers of price variation and evaluate the effectiveness of linear regression models in predicting housing values. A combination of exploratory data analysis and statistical modeling is used to gain insights into how specific property features influence market outcomes

Data & Results

The dataset includes variables such as square footage, number of bedrooms and bathrooms, lot size, year built, and neighborhood indicators. After initial cleaning and preprocessing to address missing data and skewness in price distributions, several transformations were applied to improve model interpretability. In particular, log-transformations were used to stabilize variance and linearize nonlinear relationships. Descriptive statistics and visualizations show a strong positive relationship between square footage and housing prices. Lot size is also an important determinant, though with more modest effect size. Categorical factors like neighborhood and year built exhibit notable variation in average prices across groups. The data reveal some multicollinearity, which is addressed in later model specification steps.

A multiple linear regression model was constructed with log(price) as the dependent variable. Independent variables included log(sqft), number of bathrooms, lot size, year built, and others. Model diagnostics-such as residual plots, R-squared, AIC, and VIF-were used to assess fit and address specification issues. Stepwise selection methods helped refine the model to retain relevant predictors and eliminate redundancy. The final model highlights square footage as the most significant predictor of housing prices, with an estimated elasticity close to one. Lot size and number of bathrooms are also statistically significant, while the number of bedrooms loses significance once square footage is controlled for. Newer homes typically command a premium, although this varies across neighborhoods. The regression explains a large proportion of the variance in price and satisfies key assumptions of linear modeling following transformation.

In addition to the linear model, a Random Forest model was implemented using the same dataset. The forest, built with 500 trees, explained approximately 90.46% of the variance in sale prices-outperforming the linear regression model's R-squared. This highlights the advantage of non-linear models in capturing complex patterns in housing data. However, the trade-off is reduced interpretability, which makes the OLS model still valuable for policy and economic communication

Possible Policy Implications

The policy implications of this analysis are several. First, the results suggest that housing supply constraints-such as limits on square footage expansion or density-may have direct effects on affordability. Policies that encourage more efficient land use or relax rigid zoning constraints could moderate price growth in constrained markets. Second, for developers and planners, the analysis implies that layout and design may be more important than simple bedroom count. Finally, municipalities considering land-use regulation should evaluate the extent to which local zoning laws impact market pricing dynamics.

Overall, this project demonstrates the utility of linear regression models in analyzing price variation in residential real estate. It also shows that a careful selection of variables and proper model diagnostics

are essential for drawing robust conclusions. Future research could expand the model to include spatial dependencies, interaction effects, or more nonlinear techniques to improve forecasting accuracy and better capture market complexity.