Regression Analysis of Factors Affecting House Sale Prices

Evaluating Data Quality and Key Variables Influencing Residential Property Prices

Abdul Kader

2025

Contents

[Objective 2](#_Toc201699799)

[Methodology 2](#_Toc201699800)

[Findings 2](#_Toc201699801)

[Interpretation of Model Outputs 3](#_Toc201699802)

[Insights 3](#_Toc201699803)

# Objective

The primary objective of this study is to assess the impact of various factors on the sale price of homes, using a regression model to predict the sale price based on key independent variables. The factors under consideration include Overall Quality, Living Area, Garage Capacity, and Basement Size. The purpose of this analysis is to examine the correlations between these variables and the sale price, thereby aiding in the development of pricing strategies and conducting market analysis.

# Methodology

The methodology adopted for this analysis is based on multiple linear regression, with SalePrice as the dependent variable. The independent variables selected for the model include:

**OverallQual (Overall Quality)**

**GrLivArea (Above Ground Living Area)**

**GarageCars (Number of Cars in the Garage)**

**TotalBsmtSF (Total Basement Area)**

Data from a housing dataset was analyzed using Python’s libraries, including Pandas for data manipulation and statsmodels for regression analysis. Data preprocessing included handling missing values, removing outliers, and transforming skewed data (log transformation for SalePrice). A stepwise approach was used to identify the most significant predictors.

Regression equation after log transformation of SalesPrice:

# Findings

The regression model revealed the following key findings:

* **OverallQual** has the highest positive impact on **SalePrice**, suggesting that the overall quality of the house is the most significant determinant of its value.
* **GrLivArea** also showed a strong positive correlation with SalePrice, indicating that larger homes command higher prices.
* **GarageCars** and **TotalBsmtSF** have a moderate but significant influence on price, with both showing positive coefficients.
* The adjusted R-squared value of 0.745 indicates a good model fit, explaining 74.5% of the variance in sale prices. The F-statistic of 1021.0 (p-value: < 0.05) confirms that the model is statistically significant.
* The VIF values were all within acceptable limits, confirming no multicollinearity issues.

# Interpretation of Model Outputs

The model outputs indicate that all independent variables significantly contribute to predicting SalePrice. The coefficients demonstrate that for each unit increase in OverallQual, the SalePrice increases by approximately $5,000. An increase of 100 square feet in GrLivArea corresponds to an approximate \$2,000 increase in SalePrice.

The residual analysis reveals a minor deviation from normality in the residuals, as evidenced by the Jarque-Bera test. This suggests that, despite the model fitting well, additional refinements such as transformations or incorporating interaction terms could enhance prediction accuracy.

# Insights

* **Strategic Pricing:** Based on the model, improving the overall quality (OverallQual) of a home, such as renovations or modern updates, can significantly increase the sale price. Sellers can leverage this insight to target high-quality home improvements.
* **Market Size:** Larger homes (GrLivArea) tend to sell for higher prices, emphasizing the demand for spacious properties in the market.
* **Investment in Garage and Basement:** Homes with more garage capacity (GarageCars) and larger basements (TotalBsmtSF) have a moderately higher sale price, suggesting these features are important to potential buyers.
* **Model Refinement:** Future work could include addressing the non-normality of residuals and incorporating more complex variables such as neighborhood, proximity to amenities, and market trends.