**AMES HOUSING ANALYSIS AND PREDICTION USING MULTIPLE LINEAR REGRESSION MODEL**

by

**GOUTHAM BOMMU**

**SUNWOO LEE**

**SRIKANTH GUTTIKONDA  
JIANGUO LI**

Project report submitted to the Faculty of the Graduate School of the



in fulfillment of the requirements for ANLY 502

Supervised by: Dr. Scott Layden,

Summer 2024

Table of Contents

1. Introduction .................................................................................................................... 3

2. EDA ................................................................................................................................. 3

2.1 Univariate Analysis .............................................................................................. 3

2.1 Bivariate Analysis ................................................................................................. 5

3. Inference Analysis .......................................................................................................... 8

4. The Best Model ............................................................................................................... 9

5. Prediction ...................................................................................................................... 10

6. Conclusion .................................................................................................................... 12

**1. Introduction**

This project utilizes the Ames Housing dataset, which includes detailed information on housing sales in Ames, Iowa. The goal is to identify significant predictors of house prices and develop a robust linear regression model to predict the sale price of houses based on these predictors. By understanding the relationships between various housing attributes and their prices, we can make informed decisions in the real estate market and improve predictive accuracy for future housing sales.

**2. EDA**

**2.1 Univariate Analysis**

The univariate and bivariate analyses provide valuable insights into the distribution and relationships between house prices and other key variables. The SalePrice variable exhibits a right-skewed distribution, indicating that most houses are priced in the lower range. The Year Built and Gr Liv Area variables also show right-skewed distributions. Bivariate analyses reveal that higher Gr Liv Area and better Overall Condition are associated with higher sale prices. Additionally, different Building Types and House Styles have varying impacts on house prices, as evidenced by the box plots.

**Interpretation:**

* The SalePrice variable exhibits a right-skewed distribution, with most houses priced between $100,000 and $200,000.
* The mean sale price is higher than the median, indicating a few very high-priced houses.

A graph of a graph

Description automatically generated

**Interpretation of Summary Statistics:**

SalePrice ranges from $12,789 to $755,000, with a median of $160,000 and a mean of $180,796, indicating a right-skewed distribution.

Year Built ranges from 1872 to 2010, with a median of 1973, suggesting that the dataset includes houses from a wide range of periods, but the majority were built in the mid-to-late 20th century.

Gr Liv Area (Total Living Area) ranges from 334 to 5642 square feet, with a median of 1442 square feet and a mean of 1500 square feet, indicating a slight skew towards larger homes.

**2.2 Bivariate Analysis**

The scatter plot of SalePrice versus Year Built shows a positive relationship, indicating that newer houses tend to have higher sale prices. The trend line (in red) reinforces this positive correlation.

**A graph showing the growth of a building

Description automatically generated with medium confidence**

The scatter plot of SalePrice versus Gr Liv Area (Total Living Area) shows a strong positive relationship. As the living area increases, the sale price also tends to increase. The trend line (in red) highlights this positive correlation, indicating that larger homes generally have higher sale prices.

**A graph showing a line and a dotted line

Description automatically generated with medium confidence**

The boxplot of SalePrice by Overall Cond indicates that houses with better overall conditions generally have higher sale prices. As the overall condition rating increases from 1 to 9, there is a noticeable increase in the median sale price, although there is some variability within each condition category.

A graph with a bar graph

Description automatically generated with medium confidence

Most of the houses in the Market are in condition 5 as per box plot between House Price and Over all Condition.

A graph of a bar graph

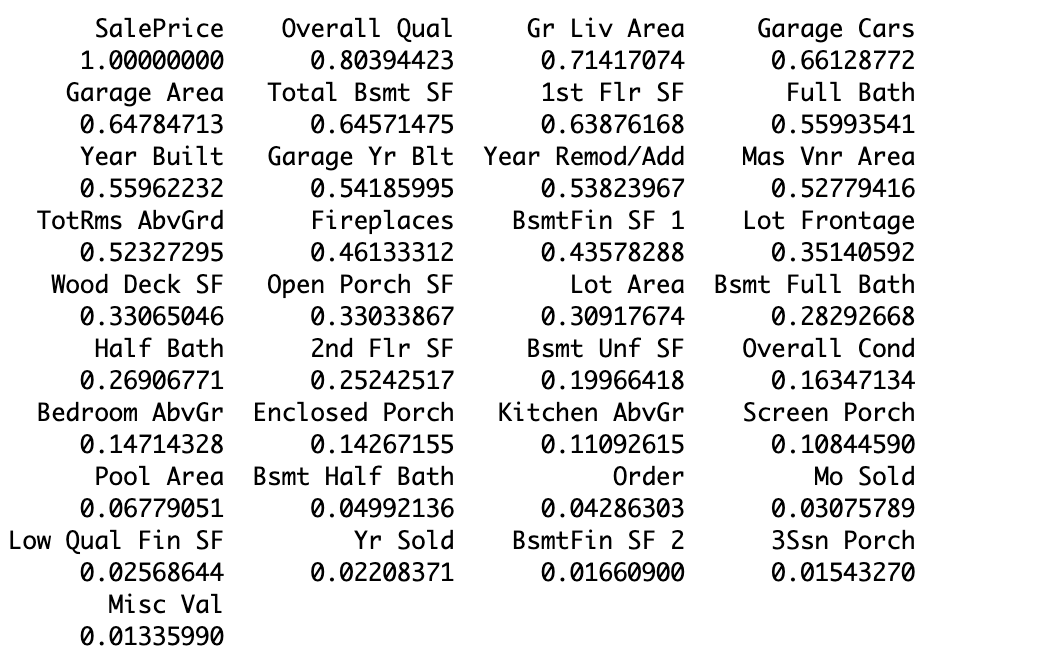
Description automatically generated

The boxplot of SalePrice by Bldg Type shows that single-family homes (1Fam) tend to have the highest sale prices, with a larger spread and more outliers compared to other building types. Two-family conversions (2fmCon), duplexes (Duplex), and townhouses (Twnhs and TwnhsE) generally have lower sale prices with less variability.

From the correlation matrix we can see these are the top 7 Predictors:

A diagram of a number of numbers

Description automatically generated with medium confidence



**3. Inference Analysis**

We did perform a multiple linear regression analysis to determine which variables significantly affect the house prices and use confidence intervals to assess the significance of the predictors.

**Research Question:**

Does the overall condition of the house significantly impact the sale price?

**Hypothesis:**

**Null Hypothesis (H0):** The overall condition of the house does not impact the sale price (βOverall Cond=0)

**Alternative Hypothesis (H1):** The overall condition of the house impacts the sale price (βOverall Cond≠0).

**Methodology:**

We conducted a multiple linear regression analysis with the log-transformed SalePrice as the response variable and included Overall Cond as one of the explanatory variables.

**Results:**

Estimate for Overall Cond: 0.0599

Standard Error for Overall Cond: 0.0029

t-value: 20.740

p-value: < 2e-16

The p-value is significantly less than 0.05, allowing us to reject the null hypothesis. This indicates that the overall condition of the house has a statistically significant impact on the sale price.

**4. The Best Model**

Using stepwise selection, the final model includes the following predictors:

Year Built, Bldg Type, House Style, Overall Cond, Overall Qual, Gr Liv Area, Garage Cars, Garage Area, Total Bsmt SF.

These variables were chosen because they showed strong correlations with SalePrice and provided meaningful insights into the factors influencing house prices.

To address issues of non-normality and heteroscedasticity, a log transformation was applied to the SalePrice variable. This transformation helps stabilize the variance and make the distribution of residuals more normal, leading to more reliable estimates.

**Model Performance:**

The final model has an Adjusted R-squared value of 0.8552, indicating that approximately 85.52% of the variance in house prices is explained by the model. The F-statistic is highly significant (p-value < 2.2e-16), suggesting that the model is a good fit.

**Significant Predictors:**

**Year Built**: Newer houses tend to have higher sale prices.

**Overall Qual**: Better quality ratings are associated with higher sale prices.

**Gr Liv Area**: Larger living areas lead to higher sale prices.

**Overall Cond**: Better overall condition ratings result in higher sale prices.

**Conditions and Implications:**

**Linearity**: The relationship between predictors and the response variable should be linear.

**Independence**: Observations should be independent.

**Homoscedasticity**: The residuals should have constant variance.

**Normality**: The residuals should be normally distributed.

Violations of these conditions can lead to biased or inefficient estimates.

**5. Prediction**

Using the final model, we can predict future house prices. The model's RMSE is 0.29602, and the MAE is 0.19071, indicating good predictive accuracy.

A screenshot of a white table

Description automatically generated

A graph showing a line of sales

Description automatically generated with medium confidence

**Prediction Equation:**

The equation for the final model, with the log-transformed SalePrice, is:

**log(SalePrice)=β0​+β1​(Year Built)+β2​(Bldg Type)+β3​(House Style)+β4​(Overall Cond)+β5​(Overall Qual)+β6​(Gr Liv Area)+β7​(Garage Cars)+β8​(Garage Area)+β9​(Total Bsmt SF)**

Where each β represents the coefficient for the corresponding predictor.

**6. Conclusion**

The analysis reveals that significant predictors of house prices in Ames, Iowa, include the house's overall quality, living area, year built, overall condition, and garage features. The model explains a substantial portion of the variance in house prices, providing a reliable tool for predicting future prices. These insights can be valuable for stakeholders in the real estate market, including buyers, sellers, and real estate agents.

**How well can you predict your response variable?**

The final model, which includes log-transformed SalePrice, demonstrates strong predictive performance with an Adjusted R-squared value of 0.8552. This indicates that approximately 85.52% of the variance in house prices is explained by the model. The root mean squared error (RMSE) and mean absolute error (MAE) values further support the model's accuracy, indicating that the model provides reliable predictions for house prices in the test dataset.

**What are the caveats to your analysis?**

Several caveats should be considered in this analysis:

**Generalizability**: The findings are specific to Ames, Iowa, and may not generalize to other housing markets with different characteristics and dynamics.

**Data Quality**: The analysis assumes that the data is accurate and complete. Any errors or missing data in the original dataset could impact the results.

**Does this data set lack information that you would have liked to use?**

While the Ames Housing dataset is comprehensive, it lacks some information that could further enhance the analysis:

**Economic Indicators**: Information on local economic conditions, such as employment rates and income levels, could provide additional context for house prices.

**Interest Rates**: Data on mortgage interest rates at the time of sale would help understand the financial environment influencing house prices.

**Market Trends**: Data on market trends over time, such as housing supply and demand dynamics, could provide a broader perspective on price fluctuations.

In conclusion, the analysis provides valuable insights into the key factors influencing house prices in Ames, Iowa, and offers a robust model for predicting house prices. However, incorporating additional data and considering the identified caveats would further refine the analysis and improve predictive accuracy.