Housing Pricing Regression Analysis

Prepared for century 21 - ames

Blake Armstrong, Jack pehlke

2025

***Executive Summary:***

Main takeaways and high points. Do we include?

Analyis 2 Table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **AIC** | **Kaggle Score** |
| Simple Linear Regression | .89 | 1272 |  | .721 |
| Multiple Linear Regression | .78 | 1590 |  | .945 |
| A Third MLR Model | .93 | 1149 |  | .288 |
| Other Models … | … | … |  | … |

***Introduction:***

This analysis focuses on a dataset that contains information about residential properties in Ames, Iowa, with the primary goal of predicting sales price of homes based on various characteristics. Specifically, we are tasked with conducting two distinct analyses that not only aim to uncover relationships between the sale price and different variables but also assess the best possible predictive model for home prices.

For the first analysis, Century 21 Ames, a real estate company specializing in homes in three neighborhoods of Ames (NAmes, Edwards, BrkSide) has commissioned the analysis, the primary question is to understand how sales price is related to the square footage of a home’s living area, and whether that relationship differs by the three neighborhoods that they specialize in. Additionally, we will calculate confidence intervals for these estimates and evaluate model assumptions to ensure the validity of the results.

The second analysis expands the scope to predict sales price for homes across all neighborhoods in Ames, Iowa. The goal here is to build a predictive model for sales price, with a focus on developing three competing models using simple and multiple linear regression techniques. The simple will be an explanatory variable of our choice, the second will be sales price predicted by living area and number of full baths, and the final will be where we selected the explanatory variables to model predicted sales price. We will denerate an adjusted R2, CV Press and Kaggle Score for each of these models and clearly describe which model you feel is the best in terms of being able to predict future sale prices of homes in Ames, Iowa.

Both analyses aim to provide valuable insights for Century 21 Ames and other stakeholders in the real estate market by developing data-driven approaches for understanding home pricing dynamics and predicting future sales. Through these two analyses, we will explore the relationships between key variables and assess the effectiveness of different modeling techniques in predicting home prices.

***Data Description:***

The dataset used for this analysis is derived from the Ames Housing Dataset, which was compiled by Dean De Cock for educational purposes in the field of data science. The Ames Housing dataset is a rich resource, containing detailed information about residential properties in Ames, Iowa, and is commonly used to teach data science techniques, including regression analysis, feature engineering, and model selection.

**Dataset Overview**

* **Number of Observations:** 1,460
* **Number of Variables:** 79 explanatory variables
* **Target Variable:** **SalePrice** (the sale price of each home)

The dataset includes a wide variety of variables that describe the physical characteristics of homes, as well as some related to their location and market conditions. These features encompass information about the house's size, condition, architectural style, materials used, neighborhood, and more. Some of the variables are continuous (e.g., square footage of the living area), while others are categorical (e.g., neighborhood or garage type).

To find out more about the data, the following link has various files that fully explain the dataset. [Kaggle Link to Data](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data)

Additionally, to view and interact with the data, we have created an RShiny app that can be accessed through the following link and a sample output of the Sales Price vs. Gross Living SqFt, separated by the three initial neighborhoods in question is pictured below (or next page). [RShiny App Link](https://barms.shinyapps.io/RealEstateAnalysis/)

A graph showing a number of colored dots

AI-generated content may be incorrect.

**Specific Variables by Analysis**

* **Analysis 1**
  + Sales Price, Gross Living Area (SqFt), and Neighborhood
* **Analysis 2**
  + **Simple Linear Regression**
    - Sales Price, XXXXXX
  + **Multiple Linear Regression**
    - Sales Price, Gross Living Square Footage, and Full Bath
  + **Custom Multiple Linear Regression**
    - Sales Price, XXXXXX

***Analysis 1:***

The goal is to build a model that quantifies the relationship between SalePrice and GrLivArea, examining potential neighborhood differences. Confidence intervals will be provided for these estimates, and model assumptions will be validated. Additionally, we will identify and address any outliers or influential observations. The result will offer Century 21 Ames insights into how home size and neighborhood location influence pricing within these three neighborhoods.

1. Model Build and Fit
   1. To answer this question we built a model that includes interaction terms between living area and neighborhoods to determine if there is a difference in the slopes/intercepts of the regression lines between each neighborhood.
2. Checking Assumptions
   1. Residual Plots
      1. Plots of the residuals can be found in the appendix under “Without Outliers” analysis, the residuals pictured are satisfactory to continue with the model.
   2. Influential point analysis (Cook’s D and Leverage)
      1. The Edwards neighborhood has two extremely large houses that were built in 2008. These homes had extremely large leverage and were excluded from the analysis as there are concerns about the sales price and the housing crisis overlap causing skew. Doing this reduced the maximum Cook’s D from around 3 to 0.3.
   3. Assumptions
      1. Linear Trend: The data does appear to be linear correlated
      2. Residuals: Addressed above in 2.a
      3. Variance: Does not appear to be any obvious violations of homoscedasticity.

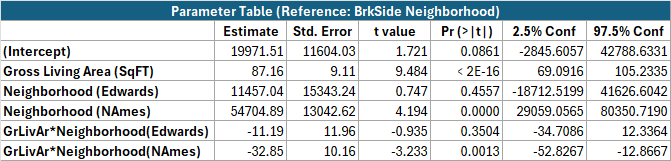
Comparing Competing Models?

Adj R2

Internal CV Press

AIC

1. Parameters
   1. Estimates



* 1. Interpretation

The BrkSide and Edwards neighborhood are statistically similar as the reference is BrkSide and both the standalone parameter and interaction parameter of Edwards in relation to BrkSide is firmly not statistically significant (p-value: 0.46 & 0.35 respectively). Meanwhile the NAmes neighborhood adjustment parameters are statistically significant and will be modelled separately.

1. Conclusion

From the analysis we can conclude that the estimated mean increase of sales price in both BrkSide & Edwards neighborhoods is statistically similar. Furthermore, it is estimated that for every 100sqft increase in gross living area, that the associated mean sales price will increase by $8,716. We are 95% confident that the true mean increase in sales price is between ($6,909, $10,523) per 100sqft increase.

Conversely the NAmes neighborhood does differ statistically from the other two in question and mean sales price is estimated to increase by $5,431 per 100sqft increase. We are 95% confident that the true increase in mean sales price lies between ($1,626, $9,236).

***Analysis 2:***

In this analysis, the goal is to build the most predictive model for home sale prices in Ames, Iowa, using all available neighborhoods in the dataset. Specifically, we will compare three competing models:

1. A simple linear regression model with a single explanatory variable chosen by the team.
2. A multiple linear regression model using GrLivArea and FullBath as predictors.
3. A multiple linear regression model using **XXXX and XXXX and XXXX** as predictors, which were chosen by the team.
4. *SLR Model*
   1. Model Build and Fit
   2. Checking Assumptions
   3. Residual Plots
   4. Influential point analysis (Cook’s D and Leverage)
   5. Assumptions
5. *MLR 1*
   1. Model Build and Fit
   2. Checking Assumptions
   3. Residual Plots
   4. Influential point analysis (Cook’s D and Leverage)
   5. Make sure to address each assumption.

1. *MLR 2*
   1. Model Build and Fit
   2. Checking Assumptions
   3. Residual Plots
   4. Influential point analysis (Cook’s D and Leverage)
   5. Assumptions

*Comparing Competing Models*

* 1. Adj R2
  2. Internal CV Press
  3. Kaggle Score
  4. AIC

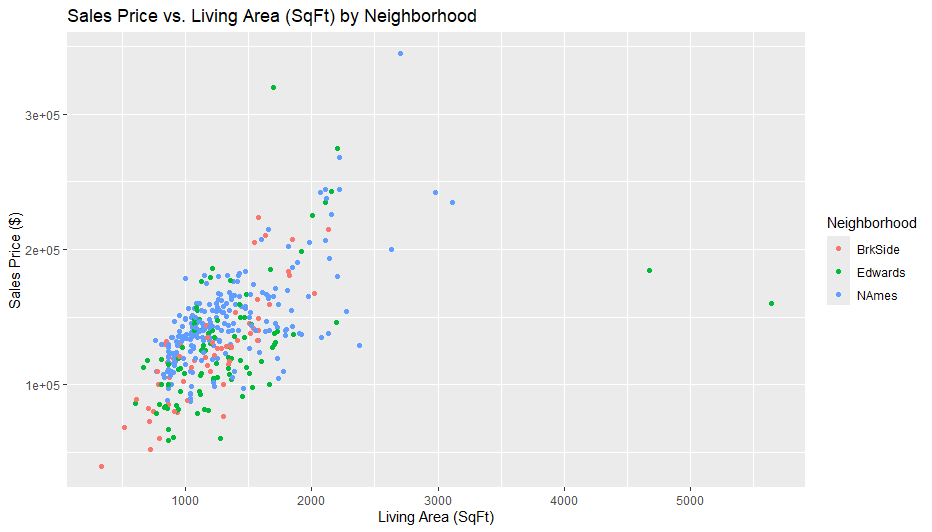
*Conclusion:*

A short summary of the analysis.

***Appendix:***

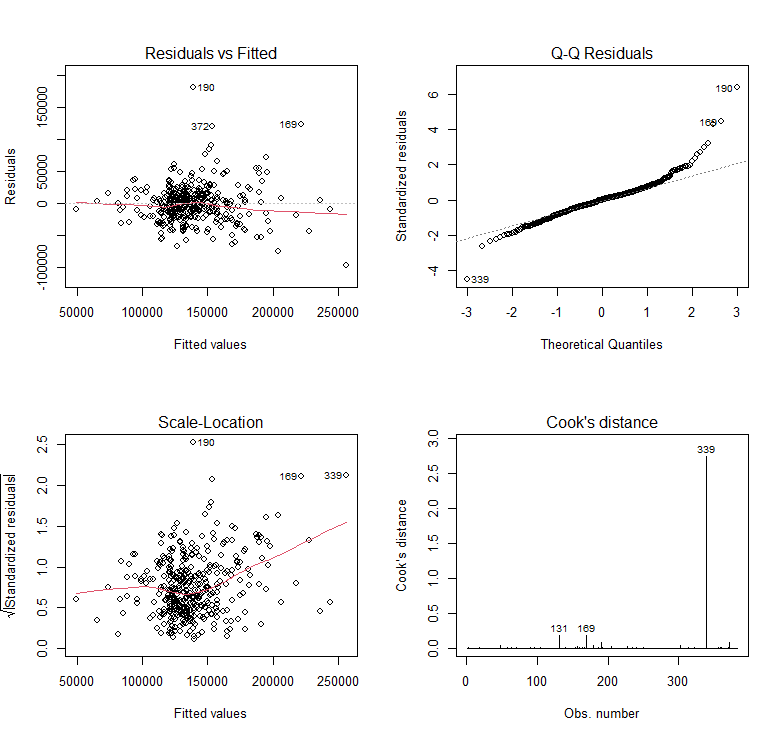
***Analysis 1:***

With Outliers:

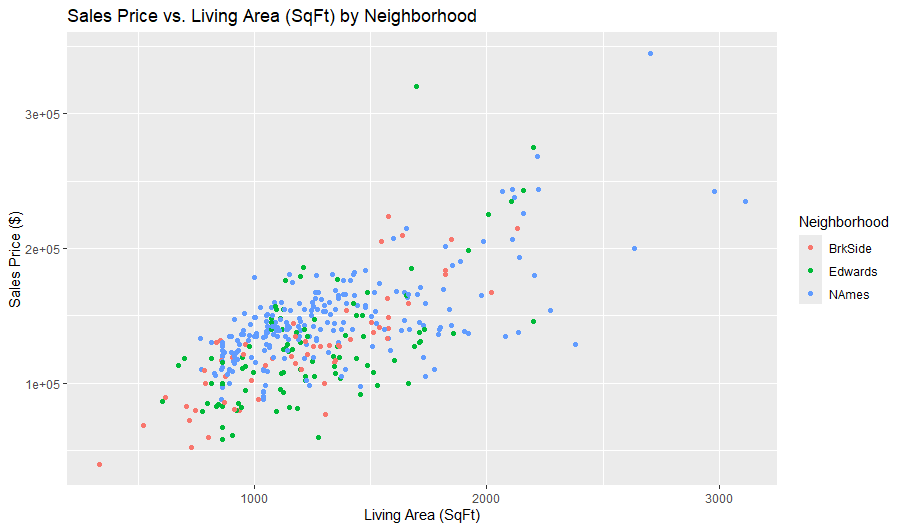


A screenshot of a computer

AI-generated content may be incorrect.

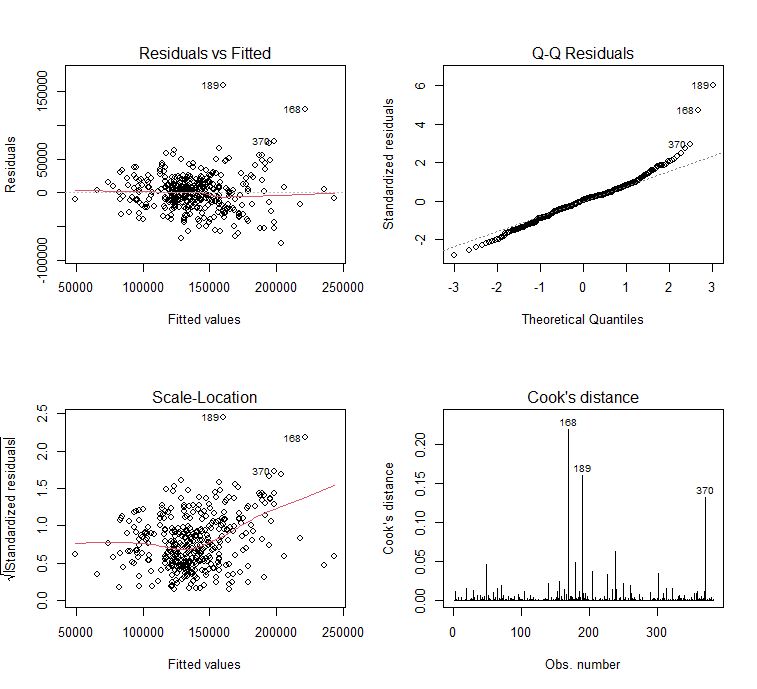


Without Outliers:



A screenshot of a computer

AI-generated content may be incorrect.



Analysis 2:

All Code:

Rubric

Presentation/Formatting (30%):

Consistent formatting / style

Organized paper with title, headings, subheadings, etc.

Labeled plots, figures, tables and charts.

Every plot, figure, table and chart included is referenced in the paper and vice versa.

No spelling or grammatical errors.

Analysis Question 1: (35%)

Clear interpretations

Correct interpretations / Useful model

Thorough interpretations

Analysis Question 2: (35 %)

Following the directions

Checking Assumptions

Comparing models / Complete Evaluation table