Modeling How Living Area Impacts Home Prices Across Neighborhoods

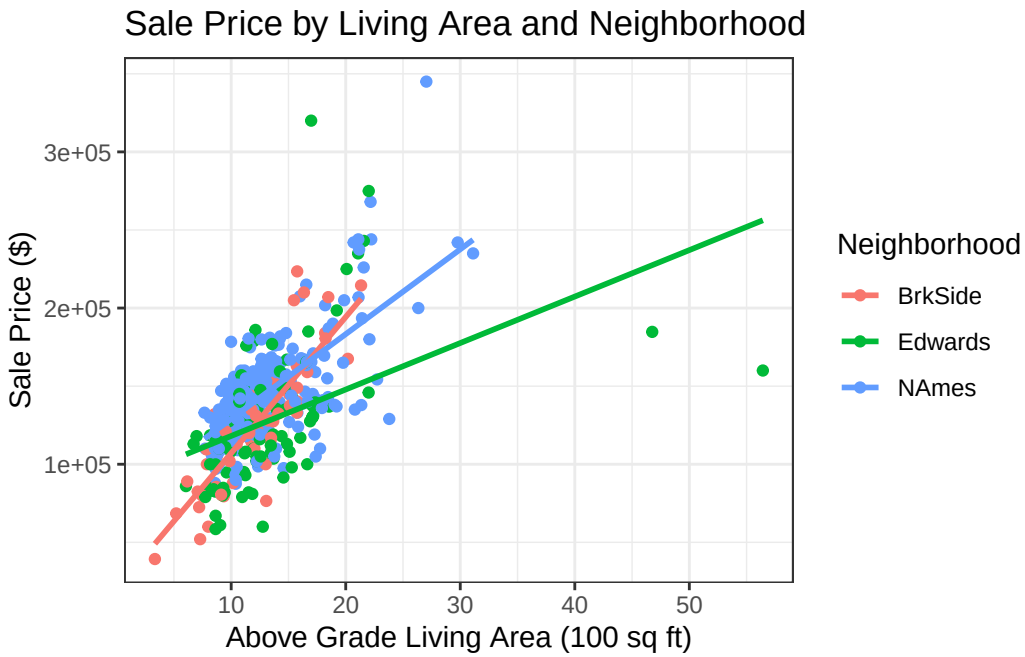
Katelyn McDonald

Table of contents

Problem Statement

Century 21 Ames is interested in understanding how the sale price of homes in three neighborhoods, Brookside, Edwards, and North Ames, is related to the size of the home's above grade living area. Because realtors usually discuss square footage in increments of 100 square feet, the analysis uses living area measured in 100 square feet units. The goal of this project is to quantify how much additional living area contributes to higher sale prices and to determine whether this relationship differs across neighborhoods. The output will be a model for the agency of sale price based on neighborhood and living area.

A scatterplot of sale price versus living area, colored by neighborhood, displays the differing price patterns across the three neighborhoods. The fitted regression lines illustrate that Brookside line has the steepest slope, while both Edwards and North Ames show shallower increases with living area.



Model Selection

The first model fit was where the neighborhoods differ only by intercept which implies that the added value of extra living area is the same across all three locations. The second model included an interaction between neighborhood and Gr100, allowing each neighborhood to have its own slope. Comparing the two models, the model with the interaction term has all predictors are significant at the .001 level excluding the intercept, while the base model has neighboord = Edwards as not statistically significant. The R^2 also drastically improved from 0.397 to 0.447 with the interaction model. This interaction model is written as:

$$\begin{aligned}\text{SalePrice}_i = \beta_0 + \beta_1 \cdot \text{Edwards}_i + \beta_2 \cdot \text{NAmes}_i \\ + \beta_3 \cdot \text{Gr100}_i + \beta_4 \cdot (\text{Edwards}_i \times \text{Gr100}_i) + \beta_5 \cdot (\text{NAmes}_i \times \text{Gr100}_i) + \varepsilon_i\end{aligned}$$

where

$$\text{Edwards}_i = 1 \quad \text{if the home is in Edwards, 0 otherwise}$$

$$\text{NAmes}_i = 1 \quad \text{if the home is in North Ames, 0 otherwise}$$

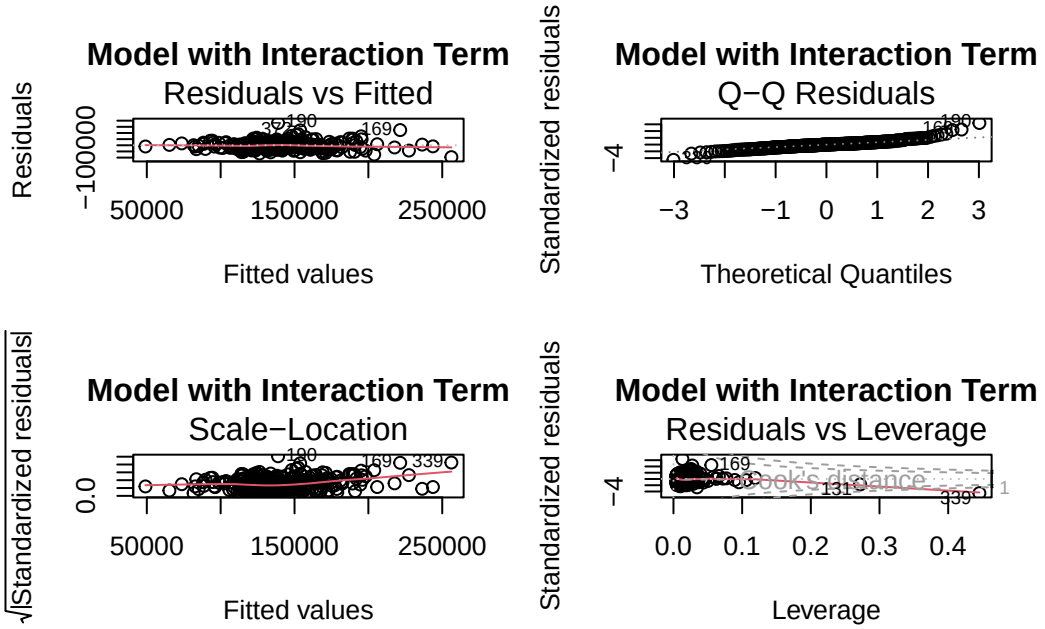
$$\text{Gr100}_i = \text{GrLivArea}_i/100$$

Addressing Diagnostics

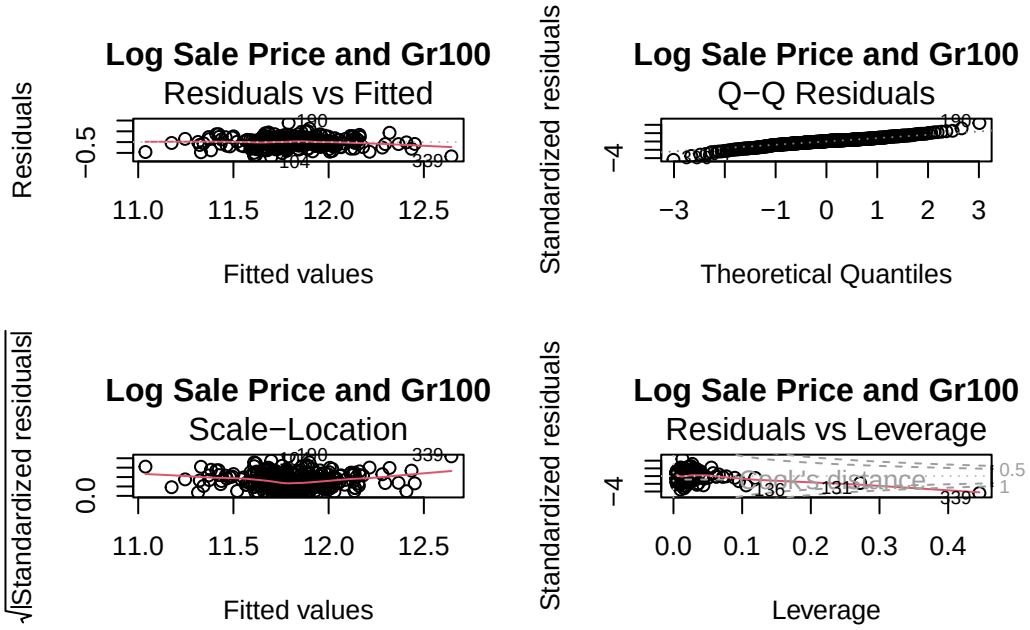
The model with the interaction terms using SalePrice on the original scale, a log-linear model using log(SalePrice), and a log(SalePrice)–log(GrLivArea) model were all compared for diagnostics. Although the log–log model produced the highest adjusted R^2, this is less suitable for the task at hand because this improvement came at the cost of interpretability. The coefficient on living area represented an elasticity rather than a dollar or percent based effect tied directly to 100 square foot increments. Since the client specifically prefers interpretation in units of 100 square feet, the log–log model was not selected.

The primary comparison was then between the interaction model and the log-linear model. Both models had statistically significant coefficients for neighborhood, living area, and the interaction terms. However, the diagnostic plots had significant improvement from the log-linear model.

Diagonsitcs plots will be reproduced to demonstrate how well each model satisfied the regression assumptions. With the interaction model, the residuals vs fitted plot displayed a clear funnel pattern, indicating strong heteroscedasticity or unequal variance as fitted values increased. The scale–location plot also showed this expansion of variance, and the Q–Q plot showed slight deviations from normality. Several observations displayed both high leverage and elevated Cook's distances, suggesting possible influential data points. Diagnostic plots for the interaction model is shown below.



By contrast, with the log-linear model, the residuals vs fitted plot showed a more even scatter around zero. The scale–location plot was considerably flatter, indicating a more equal spread of variance across the range of fitted values. Normality also improved as the Q–Q plot was much closer to a straight line. There were still a view values displaying both high leverage and elevated Cook's distances, so these will be investigated in the next section. These improvements show that the log-linear model satisfies the assumptions of linearity, constant variance, and normality for regression assumptions and for these reasons this model will be used for the analysis. Diagnostic plots for the log-linear model is shown below.



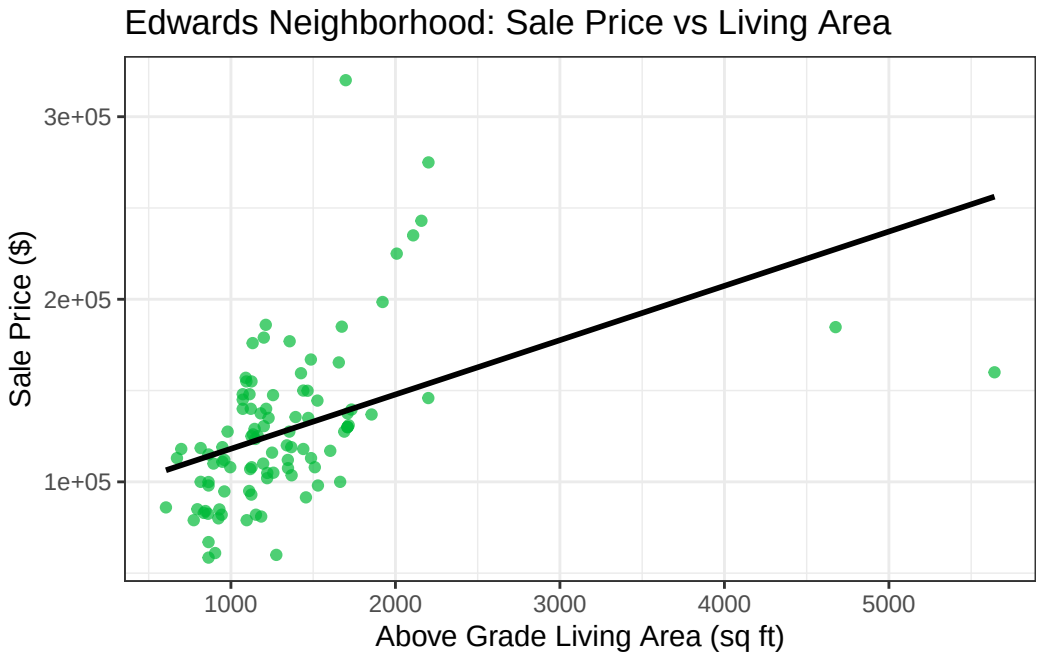
Influential Point Analysis

There is one home in the Edwards neighborhood that stands out in every diagnostic measure. This home has a Cook's distance of 2.61, a studentized residual of –4.53, and a leverage value of 0.445. With 5,642 square feet of living area, it is the largest home in Edwards. The next largest home is 4,676 square feet, and the majority of Edwards homes fall well below this range. Because this property is a legitimate part of the Edwards housing market, and because the goal of the analysis is to model all homes in the neighborhood, the point is retained despite its large influence on the fitted line.

All points with large studentized residual are within the Edwards neighborhood. These homes have unusual sale prices than what the model expects. There are two homes that sold much more than predicted, with 1698 sqft and studentized residual of 4.55 and 2201 sqft with studentized residual of 3.19. There were three homes that sold for less than what was expected by the model. One of those homes was the home discussed earlier with 5642 sqft and a studentized residual of -4.53 which sold for far less than expected given its size. Two other homes with 864 sqft (-3.21 studentized residual) and 1276 sqft (-3.53 studentized residual) also sold for less than what we expected. While these points are rare, they are not neccessarily impractible. These large residuals could have resulted from remodeling or a lack of remodeling that affected the prices, condition of the home, or even location within the neighborhood. None of the points present themself to be an error or significant deviation from what is expected so they will all be retained in the final model.

All remaining observations with large studentized residuals also come from the Edwards neighborhood. These homes have sale prices that differ more from the model's predictions than expected. Two homes sold for noticeably more than predicted. One with 1,698 square feet (studentized residual 4.55) and one with 2,201 square feet (studentized residual 3.19). Three others sold for less than expected, including the 5,642 square foot home discussed above (–4.53), as well as homes with 864 square feet (–3.21) and 1,276 square feet (–3.53). While these residuals are large, they are not unreasonable. Differences in remodeling, home condition, or location within the neighborhood could easily lead to homes selling far above or below the general trend. None of these observations appear to be data errors or implausible homes, so all are retained in the final model.

A scatterplot restricted to the Edwards neighborhood (shown below) helps illustrate the distribution of living area and sale prices among these homes. The influential points stand out, but they fall within the realistic range of values for Edwards.



Several additional homes in the full dataset have high leverage values simply because they have unusually large or unusually small living areas. High leverage alone does not make an observation problematic. It only indicates that the point has the potential to affect the slope. For most of these homes, the studentized residuals are small and Cook's distances fall well below the standard cutoff. While the leverage is high, they still align well with their neighborhood's trend. Because the neighborhoods genuinely differ in typical home size, these extreme values naturally occur and represent meaningful parts of the housing market. As a result, these points were also retained.

Final Model and Interpretations

Before fitting the final model, living area was centered at its sample mean. Centering reduces the correlation between living area and the interaction terms which improves interpertability. Before centering, neighborhood coefficients corresponded to the expected price at 0 square feet, an unrealistic and extrapolated value that increased their statistical significance. After centering, neighborhood differences are evaluated at the average home size, which provides a more meaningful comparison and explains why the neighborhood difference for Edwards was no longer significant, while all slope and interaction terms retained their significance. With the new variable Gr100.c equalling Gr100 – mean(Gr100) and Brookside as the reference neighborhood (Edwards = 0, NAmes = 0), the final model is the following:

$$\begin{aligned}\log(SalePrice) = 11.752637 &- 0.048604(Edwards) + 0.112660(NAmes) \\ &+ 0.073822(Gr100.c) - 0.052153(Edwards \times Gr100.c) - 0.041410(NAmes \times Gr100.c)\end{aligned}$$

Brookside (reference group)

For homes in Brookside, each additional 100 sq ft above the average increases the median sale price by about 7.7%. The median sale price for an average sized Brookside home is approximately \$120,452 to \$134,090.

$$log(SalePrice) = 11.7526 + 0.07382(Gr100.c)$$

Edwards

For an average sized home in Edwards, the median sale price is about 4.7% lower than Brookside, but this difference is not statistically significant because the interval includes 0. Each additional 100 sq ft increases the median price by about 2.2%, significantly less than Brookside's 7.7% due to the negative interaction term.

$$log(SalePrice) = 11.7040 + 0.02167(Gr100.c)$$

North Ames

For an average sized home in North Ames, the median sale price is about 11.9% higher than Brookside, a statistically significant difference. Each additional 100 sq ft increases the median price by about 3.3%, also smaller than Brookside's slope, reflecting the negative interaction term.

$$log(SalePrice) = 11.8653 + 0.03241(Gr100.c)$$

Overall, extra living area increases sale price in all neighborhoods, but the strength of that increase differs meaningfully across neighborhoods. The interactions highlight that Edwards and North Ames have flatter slopes, meaning additional square footage adds less value there compared to Brookside.

Conclusion

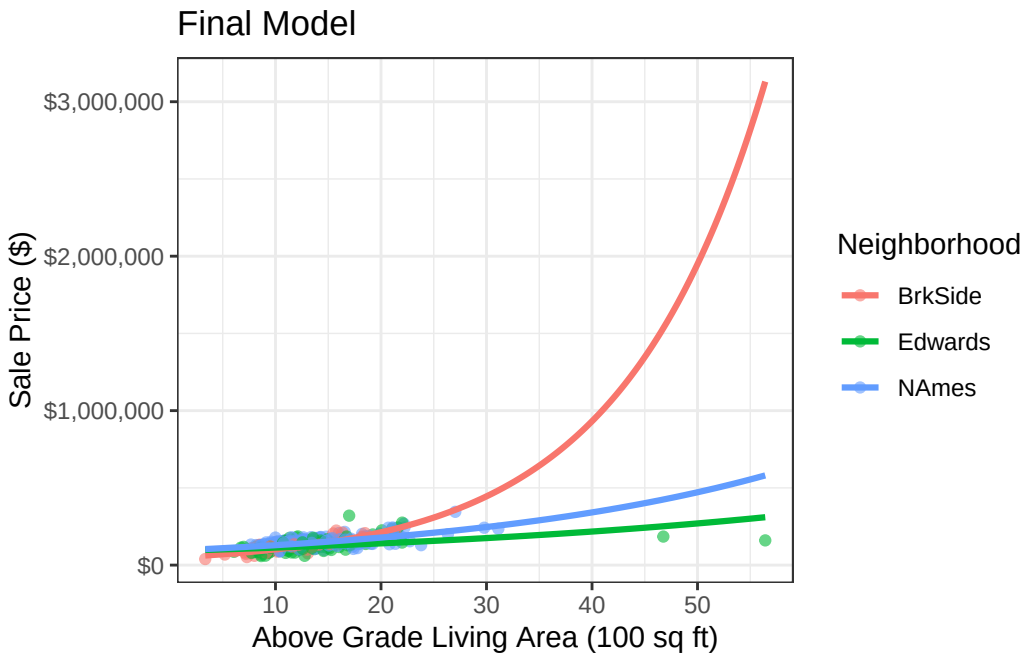
The centered log–linear interaction model shows that living area is a strong and statistically significant predictor of sale price across all three neighborhoods. However, the effect of added square footage differs considerably by location. Brookside has the steepest slope, meaning buyers in this neighborhood place the highest value on extra living space. Both Edwards and North Ames have flatter slopes, indicating that prices raise more moderately with additional square footage compared to Brookside. These differences in slope do not reflect whether a neighborhood is more or less expensive overall but reflect how each neighborhood's market responds to increased home size. A possible limitation of the model is that Brookside does not have extremely large homes in the dataset, unlike Edwards, which means the slope estimate for Brookside may be less precise for larger homes. Despite this, the model accurately reflects the general patterns of home prices in the neighborhood.

Because living area was centered, the intercepts correspond to the expected price of an average sized home in each neighborhood, and the slopes measure the percent change in median sale price for each additional 100 square feet. The model satisfies key regression assumptions, and provides interpretable estimates.

Overall, the results show that all three neighborhoods value increased living area, but Brookside values additional space most strongly. Edwards and North Ames have more moderate price growth as homes increase in size. Incorporating the interaction between neighborhood and living area is important for accurately describing the housing market and for providing pricing guidance.

Take Home Message

Home prices rise as houses get larger in all three neighborhoods, but the amount of value added by extra space isn't the same everywhere. In Brookside, buyers place the highest premium on additional square footage meaning bigger homes tend to see the strongest jump in price. In Edwards and North Ames, the starting prices of homes differ from Brookside, but extra space adds value more gradually. This means the same increase in size can lead to very different increases in price depending on the neighborhood. When advising clients, setting asking prices, or comparing homes, it is important to consider not just how big a house is, but where it is located. Neighborhood differences shape how much extra space is worth in the local market, and this model gives a clear, data-based view of those patterns. A scatterplot with the final model is produced below.



Appendix

```r
# load libraries & read in data
library(tidyverse)
library(janitor)
library(naniar)
library(ggplot2)
library(dplyr)
library(car)
library(scales)
library(MASS)

train = read.csv("/Users/katelyn/Desktop/Projects/House Price Project/data/train.csv")
glimpse(train)
```

```r
# subset data
data = train %>%
  select(Neighborhood, GrLivArea, SalePrice) %>%
  filter(Neighborhood %in% c("BrkSide", "Edwards", "NAmes")) %>%
  mutate(Neighborhood = factor(Neighborhood),
         Gr100 = GrLivArea / 100)
```

```r
# base model
model.1 = lm(SalePrice ~ Neighborhood + Gr100, data = data)
summary(model.1)
confint(model.1)

# interaction term
model.2 = lm(SalePrice ~ Neighborhood * Gr100, data = data)
summary(model.2)
confint(model.2)

AIC(model.1,model.2) # model.2 has the lower aic
```

```r
# log y
model.logy = lm(log(SalePrice) ~ Neighborhood * Gr100, data = data)
summary(model.logy)
par(mfrow=c(2,2))
plot(model.logy, , main = "Log Sale Price")

# log both
model.logboth = lm(log(SalePrice) ~ Neighborhood * log(Gr100*100), data = data)
summary(model.logboth)
par(mfrow=c(2,2))
plot(model.logboth , main = "Log Sale Price and Gr100")
```

```r
# diagnostics of best two models
par(mfrow = c(2,2))
plot(model.2, main = "Model with Interaction term")

par(mfrow = c(2,2))
plot(model.logy, main = "Log Sale Price and Gr100")
```

```r
# studentized residuals, leverage, cook's D
diagnostics = data %>%
  mutate(
    row.id = row_number(),
    student.resid = rstudent(model.logy),
    leverage = hatvalues(model.logy),
    cooksD = cooks.distance(model.logy) )

# cutoffs
n = nrow(data)
p = length(coef(model.logy))
```

```r
lev.cutoff = (2 * p) / n
cooks.cutoff = 4 / (n - p)

# diagnostics table
diagnostics = diagnostics %>%
 mutate(
    high.leverage = leverage > lev.cutoff,
    large.student = abs(student.resid) > 3,
    large.cooks = cooksD > cooks.cutoff,
    all = high.leverage | large.student | large.cooks )

diagnostics %>%
  filter(high.leverage) %>%
  arrange(desc(high.leverage)) %>%
  dplyr::select(1:8)
```

```r
# vif and centering
library(car)
vif(model.logy)
vif(model.logy, type = "predictor")

# centering
data = data %>% mutate(Gr100.c = Gr100 - mean(Gr100))

model.logy.c = lm(log(SalePrice) ~ Neighborhood * Gr100.c, data = data)
vif(model.logy.c)

summary(model.logy.c)
confint(model.logy.c)

diagnostics2 = data %>%
  mutate(
    row.id = row_number(),
    student.resid = rstudent(model.logy.c),
    leverage = hatvalues(model.logy.c),
    cooksD = cooks.distance(model.logy.c) )

# cutoffs
n = nrow(data)
p = length(coef(model.logy.c))
lev.cutoff = (2 * p) / n
cooks.cutoff = 4 / (n - p)
```

```r

# influencial points didn't change
diagnostics2 = diagnostics2 %>%
 mutate(
    high.leverage = leverage > lev.cutoff,
    large.student = abs(student.resid) > 3,
    large.cooks = cooksD > cooks.cutoff,
    all = high.leverage | large.student | large.cooks )

 diagnostics2 %>%
   filter(high.leverage) %>%
   arrange(desc(high.leverage)) %>%
   dplyr::select(1:8)


# vif after centering
vif(model.logy.c)
```

```r
# looking at ranges
range(data$Gr100[data$Neighborhood == "BrkSide"])
range(data$Gr100[data$Neighborhood == "Edwards"])
range(data$Gr100[data$Neighborhood == "NAmes"])

range(data$SalePrice[data$Neighborhood == "BrkSide"])
range(data$SalePrice[data$Neighborhood == "Edwards"])
range(data$SalePrice[data$Neighborhood == "NAmes"])
```