MATH 322: Applied Linear Regression - Final Project

Group Members: Jack Brown, Keegan Croft, Patrick Hayden, William Westerkamp

May 11, 2025

1. Objective

Research Questions

As stated in our existing project report:

The primary objective of this project is to conduct a multiple linear regression analysis on residential housing data from Ames, Iowa, to identify the most significant predictors of home sale prices. We will develop a statistically sound model for predicting sale prices and gain insight into economic and structural factors determining home valuation in the Midwest U.S. housing market. Our analysis has been guided by two research questions:

1. **What are the most important factors that influence the sale price of a home in Ames?**
    - In this part of the study, we identify the strongest predictors among a selection of numerical, ordinal, and categorical variables, such as living area, neighborhood, overall quality, and kitchen quality. By estimating the effect size and statistical significance of each predictor, we aim to construct a parsimonious model that evenly balances interpretability and predictive power.
2. **Does the effect of living area on sale price vary depending on the perceived quality of the home?**
    - Here, we explore the interaction effect between living area (Gr_Liv_Area) and overall quality (Overall_Qual). This will allow us to determine whether additional square footage adds more value in higher-quality homes compared to lower-quality ones. The question we posed allows us to test whether the marginal effect of one variable depends on another.

Modeling Objectives

Our modeling objectives include both prediction and inference. From a predictive standpoint, we are interested in building a model that could be used by homeowners, real estate agents, or appraisers to estimate fair market value based on property characteristics. From an inferential perspective, we are interested in understanding which variables have statistically significant relationships with sale price, how strong those relations are, and how they interact with each other.

To answer these questions, we will use multiple linear regression in R. Our approach includes data exploration, variable selection through forward and backward stepwise procedures, evaluation of regression assumptions (linearity, homoscedasticity, and normality), and inference using hypothesis tests and confidence intervals. The results of this analysis will not only answer our research questions but also provide an understanding of housing data through the lens of applied regression modeling.

2. Introduction

Dataset Overview

The dataset contains observations on the pricing of homes in the town of Ames, Iowa. 'SalePrice' is the response variable, meaning, it is the variable we are trying to predict.

Variables Used

As noted in our project development, there are many variables used in this project. We planned on using SalePrice, GrLivArea, OverallQual, YearBuilt, GarageCars, TotalBsmtSF, Neighborhood, FullBath, and LotArea. However, as we progressed through our analysis, we identified which variables were most important and which were not as influential as initially expected.

Descriptive Statistics

To better understand the distribution of our key variables, we examined summary statistics and frequency tables for both numerical and categorical predictors used in the analysis.

```
##    SalePrice       Gr.Liv.Area    Overall.Qual      Year.Built
##  Min.   : 12789   Min.   : 334   Min.   : 1.000   Min.   :1872
##  1st Qu.:129500   1st Qu.:1126   1st Qu.: 5.000   1st Qu.:1954
##  Median :160000   Median :1442   Median : 6.000   Median :1973
##  Mean   :180796   Mean   :1500   Mean   : 6.095   Mean   :1971
##  3rd Qu.:213500   3rd Qu.:1743   3rd Qu.: 7.000   3rd Qu.:2001
##  Max.   :755000   Max.   :5642   Max.   :10.000   Max.   :2010
```

These statistics reveal a right-skewed distribution for sale price and living area, with many homes built post-1950 and the average quality rating slightly above the mid-point of the 1–10 scale.

```
##
## 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer   SLvl
##    314     19   1481      8     24    873     83    128
```

```
##
##   Ex   Fa   Gd   Po   TA
##  205   70 1160    1 1494
```

The most common house style in the dataset is 1Story, followed by 2Story and 1.5Fin. Most homes have kitchens rated as typical (TA) or good (Gd). Very few homes fall into the lowest (Po) or highest quality (Ex) categories.

Visualizations

Distribution of Sale Price

The sale price distribution is clearly right-skewed, with most homes concentrated in the lower price ranges and a long tail extending to the higher-priced properties.

The log transformation effectively normalizes the distribution, making it more symmetric and suitable for linear regression analysis.

Figure 1: Histogram of Sale Price

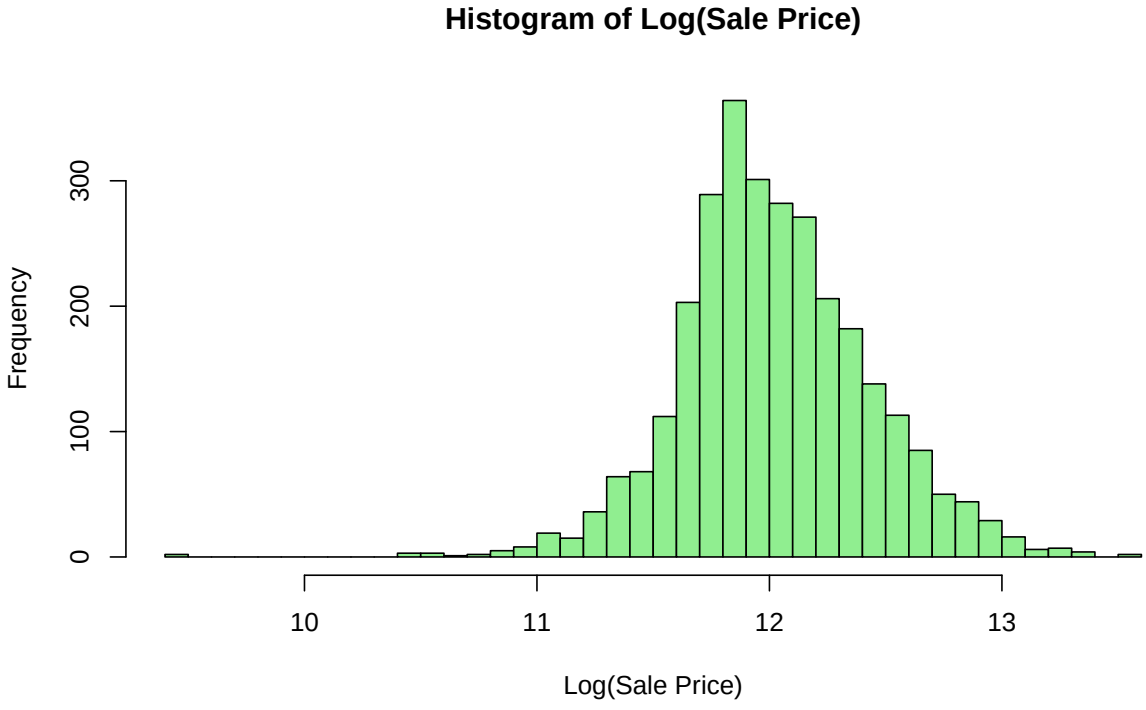


Figure 2: Histogram of Log-transformed Sale Price

Relationship Between Variables

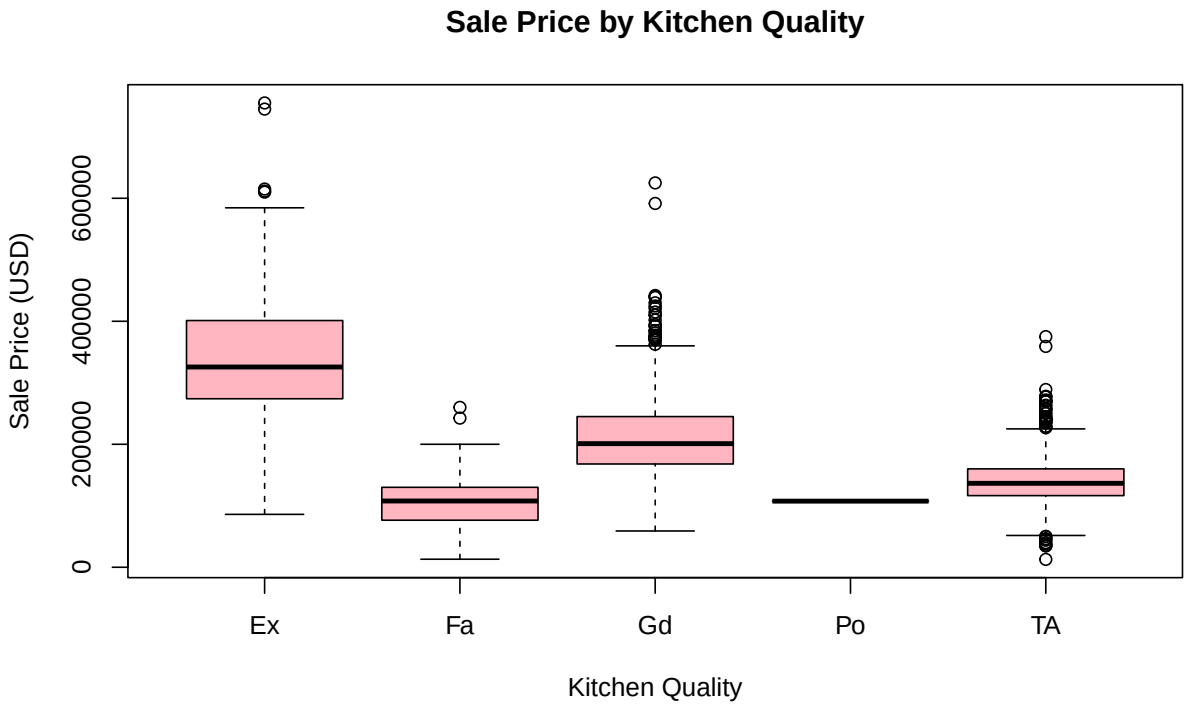


Figure 3: Sale Price by Kitchen Quality

From the boxplot, we can see that kitchen quality has a substantial impact on home prices, with higher-quality kitchens associated with significantly higher sale prices.

The scatterplot reveals a positive relationship between living area and sale price, with the points colored by overall quality showing that higher-quality homes tend to be more expensive at any given size.

3. Methodology and Analysis

Initial Model Building

To answer the question of what are the most important factors that influence the sale price of a home in Ames, we built a multiple linear regression model. We built this initial model using a curated subset of the Ames Housing dataset. The goal was to see what factors influenced the sale price of the home the most.

The sale price was log-transformed because the sale prices are very heavily right-skewed. This is because a small number of expensive homes pull the distribution to the right. The log transformation combats that by making the distribution more symmetric. We fit the model in R using the lm() function. The model included both numerical predictors and categorical predictors.

```
# Create the model (if it doesn't already exist)
if(!exists("final_model_rq1") && file.exists("data/final_model_rq1.rds")) {
  final_model_rq1 <- readRDS("data/final_model_rq1.rds")
} else {
```

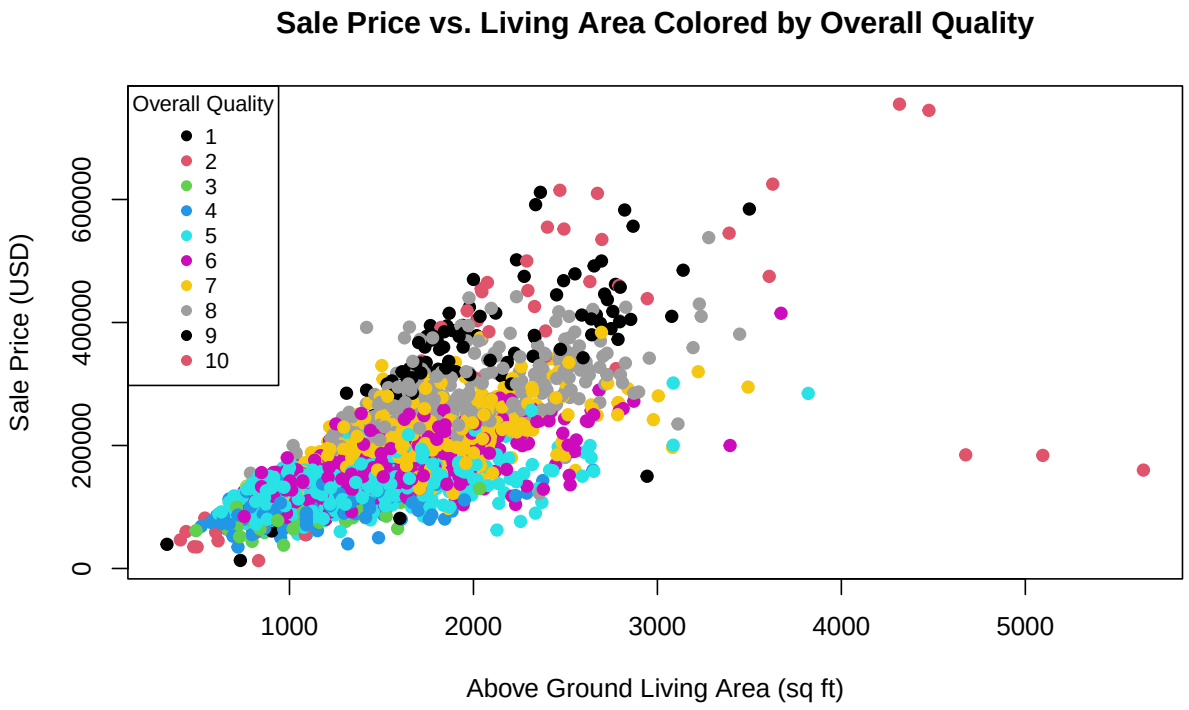


Figure 4: Sale Price vs. Living Area Colored by Overall Quality

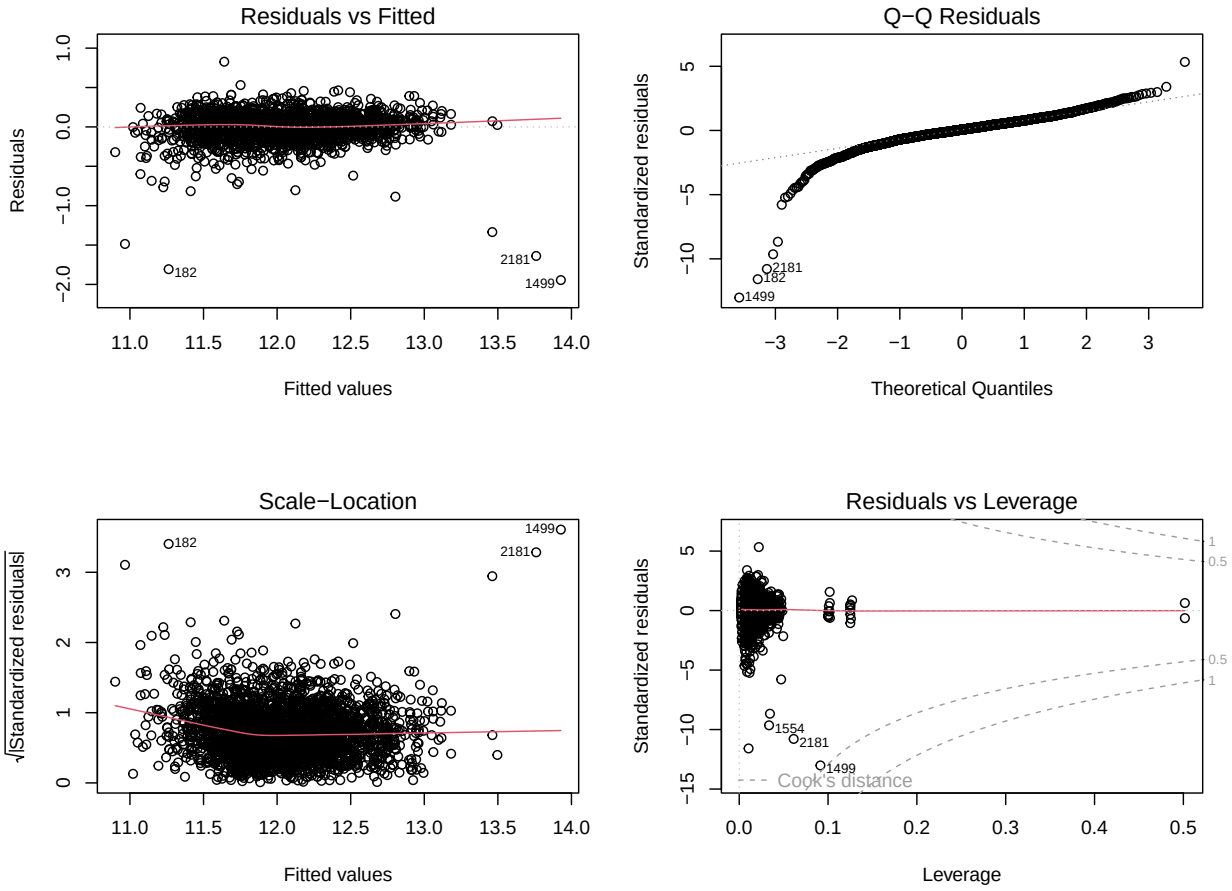
```
  # Build the model if it doesn't exist
  model <- lm(LogSalePrice ~ Gr.Liv.Area + Overall.Qual + Year.Built +
                Garage.Cars + Total.Bsmt.SF + Neighborhood + Kitchen.Qual,
              data = ames_subset)
  final_model_rq1 <- model
}

# Model summary
summary(final_model_rq1)
```

```
##
## Call:
## lm(formula = LogSalePrice ~ Gr.Liv.Area + Overall.Qual + Year.Built +
##     Garage.Cars + Total.Bsmt.SF + Neighborhood + Kitchen.Qual,
##     data = ames_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94498 -0.06697  0.00988  0.08540  0.82757
##
## Coefficients:
##                        Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)          7.952043375  0.438071424  18.152 < 0.0000000000000002 ***
## Gr.Liv.Area          0.000223569  0.000008051  27.769 < 0.0000000000000002 ***
## Overall.Qual         0.090454143  0.003806499  23.763 < 0.0000000000000002 ***
```

```
## Year.Built           0.001528550  0.000220080   6.945     0.00000000000464 ***
## Garage.Cars          0.059953161  0.005392850  11.117 < 0.0000000000000002 ***
## Total.Bsmt.SF        0.000103723  0.000008750  11.855 < 0.0000000000000002 ***
## NeighborhoodBlueste -0.060506068  0.058235096  -1.039             0.298893
## NeighborhoodBrDale  -0.174647487  0.042528144  -4.107     0.00004125994222 ***
## NeighborhoodBrkSide  0.027229269  0.036757265   0.741             0.458883
## NeighborhoodClearCr  0.208910063  0.039183549   5.332     0.00000010486794 ***
## NeighborhoodCollgCr  0.077662153  0.031278663   2.483             0.013088 *
## NeighborhoodCrawfor  0.217099263  0.035714489   6.079     0.00000000137029 ***
## NeighborhoodEdwards -0.013039498  0.033848439  -0.385             0.700094
## NeighborhoodGilbert  0.071384005  0.032529230   2.194             0.028282 *
## NeighborhoodGreens   0.053926968  0.063324485   0.852             0.394508
## NeighborhoodGrnHill  0.521322261  0.115027841   4.532     0.00000607593405 ***
## NeighborhoodIDOTRR  -0.124554422  0.037777853  -3.297             0.000989 ***
## NeighborhoodLandmrk -0.060680342  0.159794373  -0.380             0.704166
## NeighborhoodMeadowV -0.133653754  0.041166179  -3.247             0.001181 **
## NeighborhoodMitchel  0.083864921  0.033913673   2.473             0.013459 *
## NeighborhoodNAmes    0.069154878  0.032394298   2.135             0.032862 *
## NeighborhoodNoRidge  0.159958283  0.035832980   4.464     0.00000835236030 ***
## NeighborhoodNPkVill -0.042316202  0.044919127  -0.942             0.346244
## NeighborhoodNridgHt  0.144671362  0.032788168   4.412     0.00001060174427 ***
## NeighborhoodNWAmes   0.071395854  0.033534544   2.129             0.033337 *
## NeighborhoodOldTown -0.052327084  0.035822226  -1.461             0.144195
## NeighborhoodSawyer   0.073699085  0.033953620   2.171             0.030044 *
## NeighborhoodSawyerW  0.031559430  0.033170125   0.951             0.341460
## NeighborhoodSomerst  0.083241600  0.031926576   2.607             0.009173 **
## NeighborhoodStoneBr  0.163109546  0.037407032   4.360     0.00001343619636 ***
## NeighborhoodSWISU    0.016057254  0.040811848   0.393             0.694019
## NeighborhoodTimber   0.135139555  0.035101002   3.850             0.000121 ***
## NeighborhoodVeenker  0.150217862  0.044045821   3.410             0.000657 ***
## Kitchen.QualFa      -0.219057744  0.025307987  -8.656 < 0.0000000000000002 ***
## Kitchen.QualGd      -0.075585973  0.013872167  -5.449     0.00000005499067 ***
## Kitchen.QualPo      -0.363275702  0.159801471  -2.273             0.023082 *
## Kitchen.QualTA      -0.133115289  0.015711631  -8.472 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1568 on 2891 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.852
## F-statistic: 469.1 on 36 and 2891 DF,  p-value: < 0.00000000000000022
```

Model Diagnostics



The top left plot shows a mostly flat red line which is good. The top right shows that the residuals are mostly normally distributed. The bottom left plot also shows a relatively flat red line which is good. The bottom right shows a few points with high leverage but nothing out of the ordinary.

For the results, we have an adjusted R-squared of 0.852. This means that our model explains around 85% of the variance in log sale prices. All the continuous predictors were very significant. Gr.Liv.Area (t=27.8), Overall.Qual (t=23.8). So yes, the neighborhood and year built features mattered a lot, but the interior quality seems to be of utmost importance.

The most powerful predictor is by far Gr.Liv.Area, which is the square footage of the living area. Following that is overall quality, and then garage capacity. All in all, the model has excellent fit which we can see with the 0.85 R-squared.

Interaction Model for Research Question 2

For our second research question, we were looking to see if the effect of living area on sale price varies depending on the perceived quality of the home. To take a look at this, a model using only Gr.Liv.Area and Overall.Qual in relation to SalePrice was created which allowed us to see the interaction between the variables.

```r
# Create the interaction model (if it doesn't already exist)
if(!exists("interaction_model") && file.exists("data/final_model_rq2.rds")) {
  interaction_model <- readRDS("data/final_model_rq2.rds")
} else {
  # Build the interaction model if it doesn't exist
  interaction_model <- lm(log(SalePrice) ~ Gr.Liv.Area * Overall.Qual,
                          data = ames_subset)
}

# Model summary
summary(interaction_model)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ Gr.Liv.Area * Overall.Qual, data = ames_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5665 -0.1031  0.0165  0.1259  0.7520
##
## Coefficients:
##                               Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)                10.263047338  0.042625466 240.773 < 0.0000000000000002
## Gr.Liv.Area                 0.000432482  0.000028623  15.110 < 0.0000000000000002
## Overall.Qual                0.220377728  0.006901518  31.932 < 0.0000000000000002
## Gr.Liv.Area:Overall.Qual -0.000024487  0.000004065  -6.023        0.00000000192
##
## (Intercept)                ***
## Gr.Liv.Area                ***
## Overall.Qual               ***
## Gr.Liv.Area:Overall.Qual ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2 on 2924 degrees of freedom
## Multiple R-squared:  0.7594, Adjusted R-squared:  0.7591
## F-statistic:  3076 on 3 and 2924 DF,  p-value: < 0.00000000000000022
```

```r
# Compare with model without interaction
no_interaction_model <- lm(log(SalePrice) ~ Gr.Liv.Area + Overall.Qual,
                          data = ames_subset)
anova(no_interaction_model, interaction_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(SalePrice) ~ Gr.Liv.Area + Overall.Qual
## Model 2: log(SalePrice) ~ Gr.Liv.Area * Overall.Qual
##   Res.Df    RSS Df Sum of Sq      F         Pr(>F)
## 1   2925 118.39
## 2   2924 116.94  1    1.4509 36.279 0.000000001924 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary output from the interaction model has an adjusted R^2 value of 0.7594 which shows that our model is a decent fit for around 75% of the data.

A hypothesis test was performed to see if the interaction between Overall.Qual and Gr.Liv.Area was zero, and the conclusion found from it was that there was significant evidence to reject the null hypothesis and that the interaction between the variables is significant to the model. This was determined because the p-value from the ANOVA test was very small, allowing us to be confident to at least 5% that the interaction is significant.

Hypothesis test:
H : = 0 vs. H : 0
F = 36.575
p-value 0
Decision: p-value < 0.05. We reject the null hypothesis at a 5% level of significance.
Conclusion: We have significant evidence that the interaction is significant to the model.

All of this allows us to conclude that the combination of the overall quality of a house and how much square footage is available has a significant effect on the overall sale price of the house.

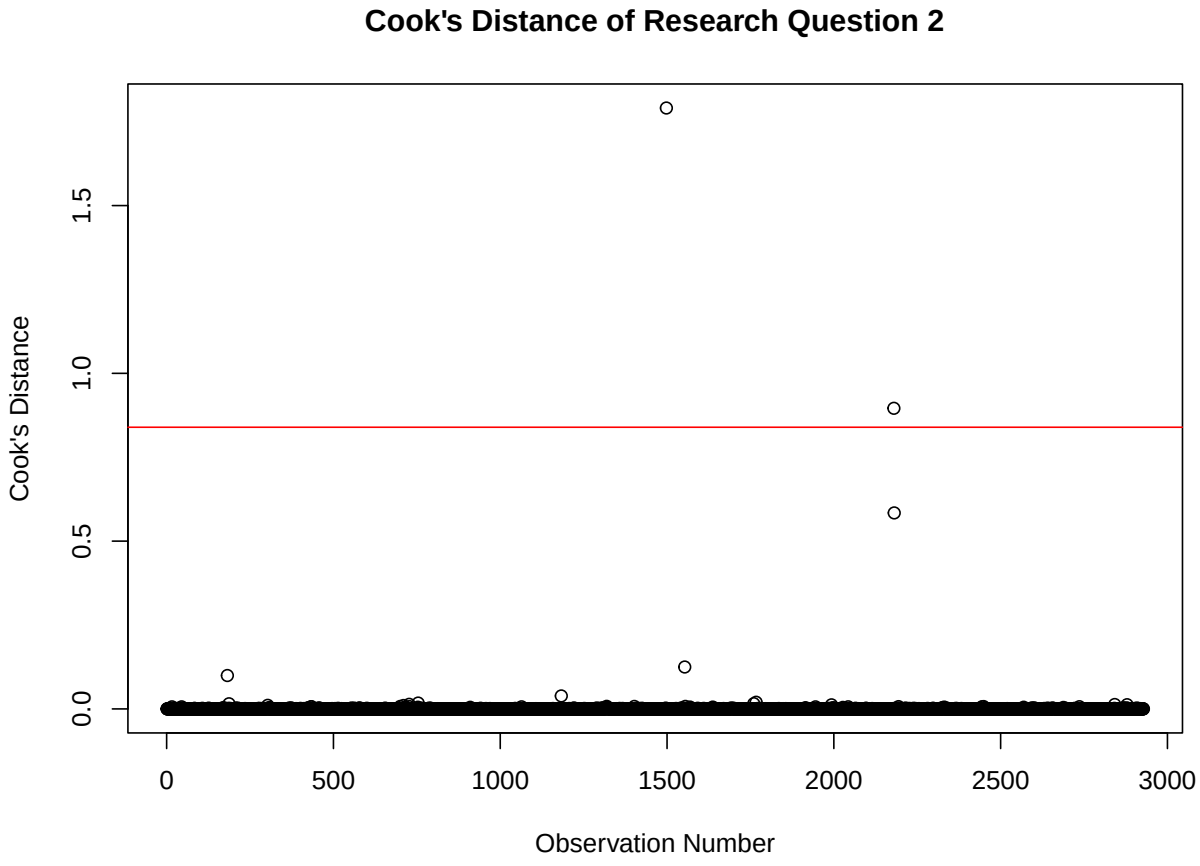


Figure 5: Cook's Distance Plot for Research Question 2 Model

This model is a better representation of what Cook's distance should be and helps identify outliers that might be affecting the data.

Practical Interpretation of Effects

```
## Interpretable effects (percentage change in price):

## - 100 sq ft increase in living area: 2.26% increase in price

## - 1 point increase in overall quality: 9.47% increase in price

## - 10 years newer home: 1.54% increase in price

## - 1 additional car garage capacity: 6.18% increase in price

## - 100 sq ft increase in basement area: 1.04% increase in price

##
## Effect of 100 sq ft increase in living area by quality level:

## - Quality level 3: 3.66% increase in price
## - Quality level 5: 3.15% increase in price
## - Quality level 7: 2.65% increase in price
## - Quality level 9: 2.14% increase in price
```

4. Results and Conclusions

Summary of Findings

Our analysis revealed several important points about the Ames housing market:

1. **Most important factors**: The most significant predictors of home sale prices:
    - Living area (Gr.Liv.Area)
    - Overall quality rating (Overall.Qual)
    - Garage capacity (Garage.Cars)
    - Basement size (Total.Bsmt.SF)
    - Year built (Year.Built)
    - Neighborhood location
    - Kitchen quality
2. **Magnitude of effects**:
    - A price rise of about 2.26% corresponds to a 100 square foot increase in living space
    - On a scale of 1 to 10, every point increase in the overall quality rating corresponds to a price increase of roughly 9.47%.
    - Prices rise by roughly 6.18% for every increase in car garage capacity
    - Homes that are ten years or younger sell for about 1.54% more
    - GrnHill and ClearCr, two upscale neighborhoods, can get prices that are more than 20% higher
3. **Interaction effect**: There is a statistically significant relationship between living area and overall quality. It's interesting to note that lower-quality homes actually have a larger percentage effect from more square footage. This implies that adding square footage may have a correspondingly greater effect on value in homes of inferior quality.

Limitations

Our analysis has several limitations:

1. **Temporal limitations**: Because the data is from a certain time period, it could not accurately represent long-term patterns or current market conditions.
2. **Geographic specificity**: Our approach is unique to Ames, Iowa, and might not translate well to other housing markets with distinct dynamics.
3. **Omitted variables**: Our model may have overlooked significant elements that influence property prices, such as lot features, school districts, or accessibility to amenities.
4. **Model assumptions**: Our diagnostic plots reveal that, despite the log transformation's assistance in mitigating right-skewness, the residuals still exhibit a few small departures from normality.
5. **Interaction complexity**: There may be more significant interactions, but we only examined one (living area $\times$ quality).

Recommendations

Based on our findings, we offer the following recommendations:

1. **For homeowners considering renovations**:
    - Give priority to raising overall quality ratings, especially those for the kitchen
    - Take into account increasing living space, as this has a big effect on home value.
    - Extending a garage might be a wise financial decision.
2. **For homebuyers**:
    - Contemplate the significant neighborhood impact on pricing
    - Assess the trade-off between size and quality of a property
    - Compute the fair market value of possible acquisitions using our model
3. **For real estate professionals**:
    - Utilize our model's percentage effects to provide clients with more insightful advice
    - Think about location-specific pricing tactics
    - When evaluating properties, take into consideration how size and quality interact
4. **For future research**:
    - Explore other interaction effects
    - Increase the number of neighborhood-level variables.
    - Examine non-linear correlations for specific predictors

By determining the primary determinants of Ames home prices and verifying that there is, in fact, an interaction effect between living area and home quality, our analysis has effectively addressed our research questions. In the Ames market, this model might be a useful instrument for valuing and making decisions about real estate.