Ames Housing Price Prediction[*]

Analysis of Structural Housing Features and Sale Prices

Andy Jiang

May 22, 2025

This paper analyzes housing sale prices in Ames, Iowa using multiple linear regression. The dataset is filtered to focus on homes with 2–4 bedrooms, built after 1940, and rated average to excellent in quality. Predictors include log-transformed ground living area, garage area, year built, number of bedrooms, and overall quality. Log transformations address skewness and heteroscedasticity. The final model achieves an R^2 of 0.819, with ground living area and overall quality identified as the most influential predictors. The results provide interpretable insights for investors, buyers, and policymakers.

Table of contents

[*] Code and data are available at: https://github.com/AndyYanxunJiang/ames-housing-price-prediction.

```r
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.2.3


Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(AmesHousing)
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

1 Introduction

This analysis investigates the relationship between structural housing features and sale price in Ames, Iowa. We use a filtered and transformed dataset and fit a multiple linear regression model to predict log-transformed sale prices.

2 Data Preparation

```r
ames_data <- make_ames()

ames_data$log_Sale_Price <- log(ames_data$Sale_Price)
ames_data$log_Gr_Liv_Area <- log(ames_data$Gr_Liv_Area)
ames_data$log_Garage_Area <- log(ames_data$Garage_Area)

valid_qual_levels <- c("Average", "Above_Average", "Good", "Very_Good", "Excellent")
```

```
valid_bedroom_levels <- c(2, 3, 4)

ames_data <- ames_data %>%
  filter(
    Overall_Qual %in% valid_qual_levels,
    Bedroom_AbvGr %in% valid_bedroom_levels,
    Year_Built >= 1940,
    Garage_Area > 0
  ) %>%
  mutate(
    Overall_Qual = factor(Overall_Qual, levels = valid_qual_levels),
    Bedroom_AbvGr = factor(Bedroom_AbvGr, levels = valid_bedroom_levels)
  )
```

3 Model Fitting

```
lm_model <- lm(
  log_Sale_Price ~ log_Gr_Liv_Area + log_Garage_Area + Year_Built + Bedroom_AbvGr + Overal
  data = ames_data
)
summary(lm_model)
```

```
Call:
lm(formula = log_Sale_Price ~ log_Gr_Liv_Area + log_Garage_Area +
    Year_Built + Bedroom_AbvGr + Overall_Qual, data = ames_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02905 -0.07421  0.00451  0.08624  0.55846

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.681316   0.458769   5.845 5.89e-09 ***
log_Gr_Liv_Area            0.532503   0.017795  29.925  < 2e-16 ***
log_Garage_Area            0.147952   0.012221  12.107  < 2e-16 ***
Year_Built                 0.002299   0.000232   9.906  < 2e-16 ***
Bedroom_AbvGr3            -0.015591   0.008003  -1.948 0.051521 .
Bedroom_AbvGr4            -0.072136   0.013378  -5.392 7.76e-08 ***
Overall_QualAbove_Average  0.032241   0.009410   3.426 0.000624 ***
```

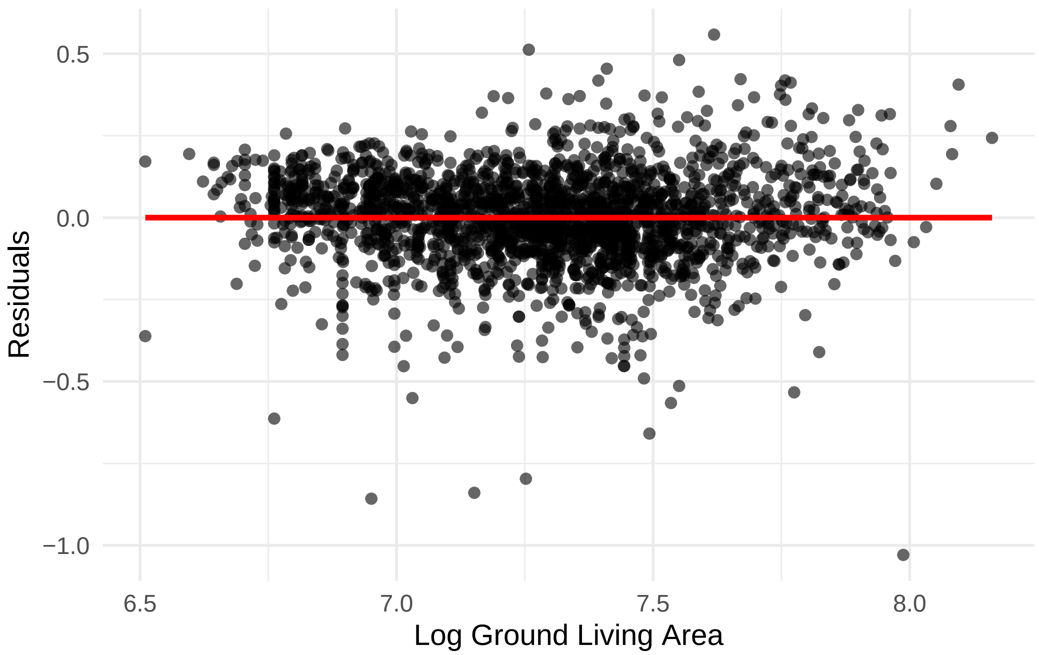
```
Overall_QualGood           0.113127   0.012178   9.289  < 2e-16 ***
Overall_QualVery_Good      0.268864   0.015140  17.759  < 2e-16 ***
Overall_QualExcellent      0.483256   0.020414  23.673  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1425 on 2064 degrees of freedom
Multiple R-squared:  0.819, Adjusted R-squared:  0.8182
F-statistic:  1038 on 9 and 2064 DF,  p-value: < 2.2e-16
```

4 Residual Diagnostics

```r
ames_data$residuals <- resid(lm_model)

# Residual plots
plot_list <- list(
  ggplot(ames_data, aes(x = log_Gr_Liv_Area, y = residuals)) +
    geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(x = "Log Ground Living Area", y = "Residuals") + theme_minimal(),

  ggplot(ames_data, aes(x = Year_Built, y = residuals)) +
    geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(x = "Year Built", y = "Residuals") + theme_minimal(),

  ggplot(ames_data, aes(x = log_Garage_Area, y = residuals)) +
    geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(x = "Log Garage Area", y = "Residuals") + theme_minimal(),

  ggplot(ames_data, aes(x = Bedroom_AbvGr, y = residuals)) +
    geom_boxplot() + labs(x = "Bedrooms Above Ground", y = "Residuals") + theme_minimal(),

  ggplot(ames_data, aes(x = Overall_Qual, y = residuals)) +
    geom_boxplot() + labs(x = "Overall Quality", y = "Residuals") + theme_minimal()
)

plot_list
```

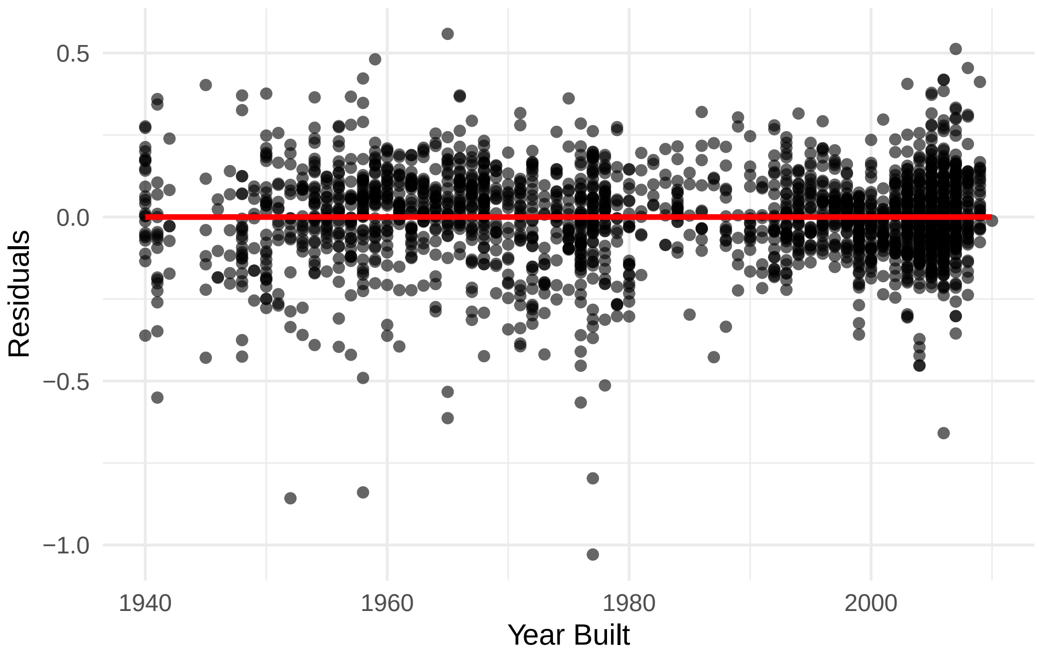
```
[[1]]
```

```
`geom_smooth()` using formula = 'y ~ x'
```



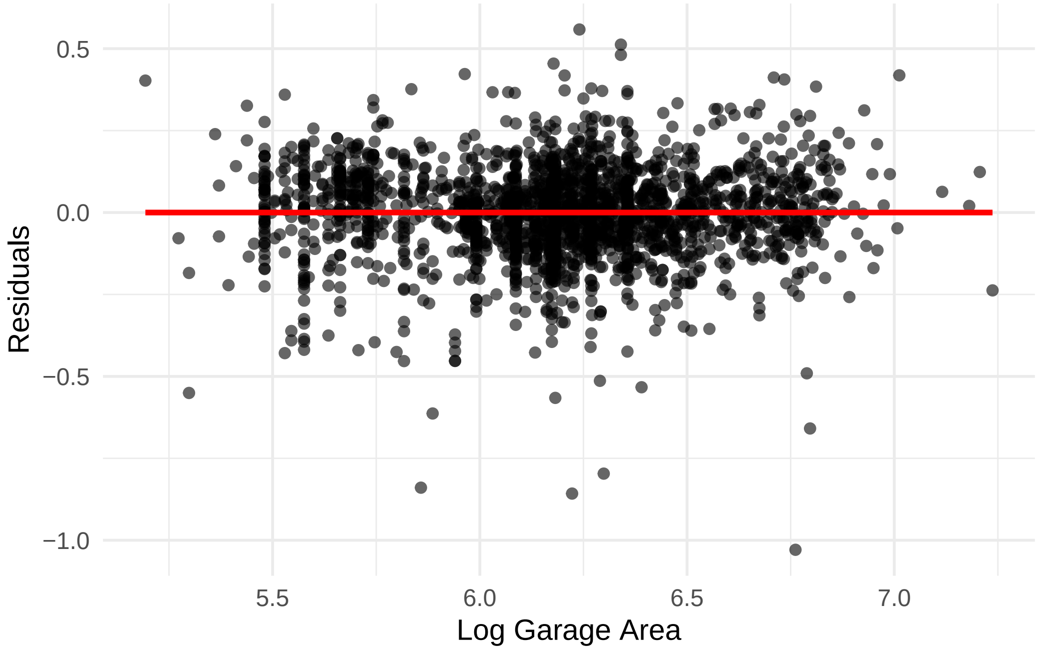
```
[[2]]
```

```
`geom_smooth()` using formula = 'y ~ x'
```

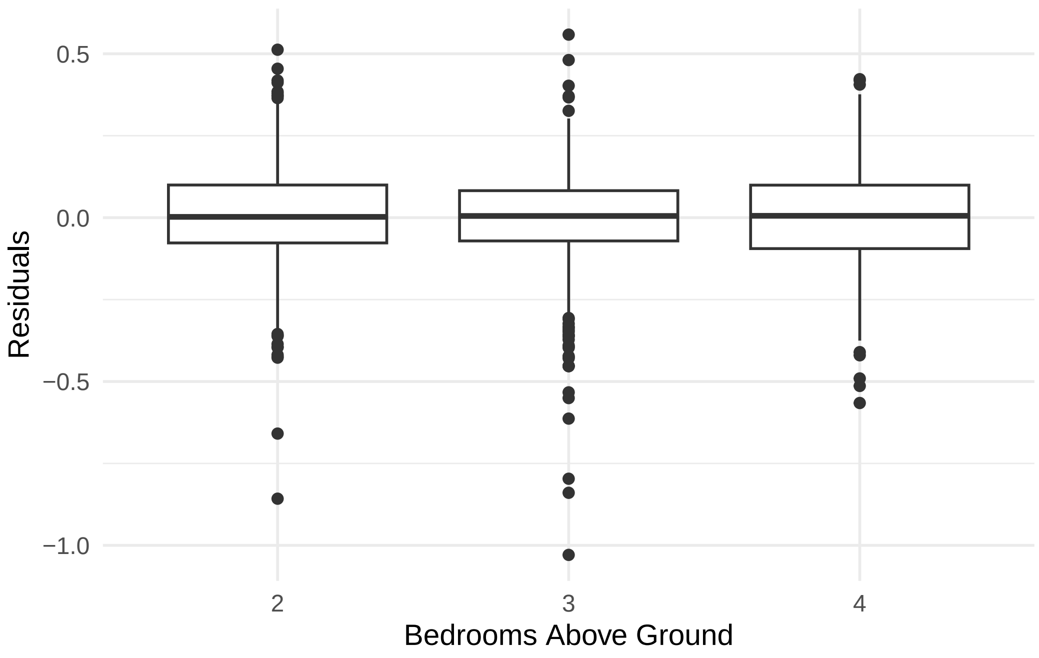


```
[[3]]
```

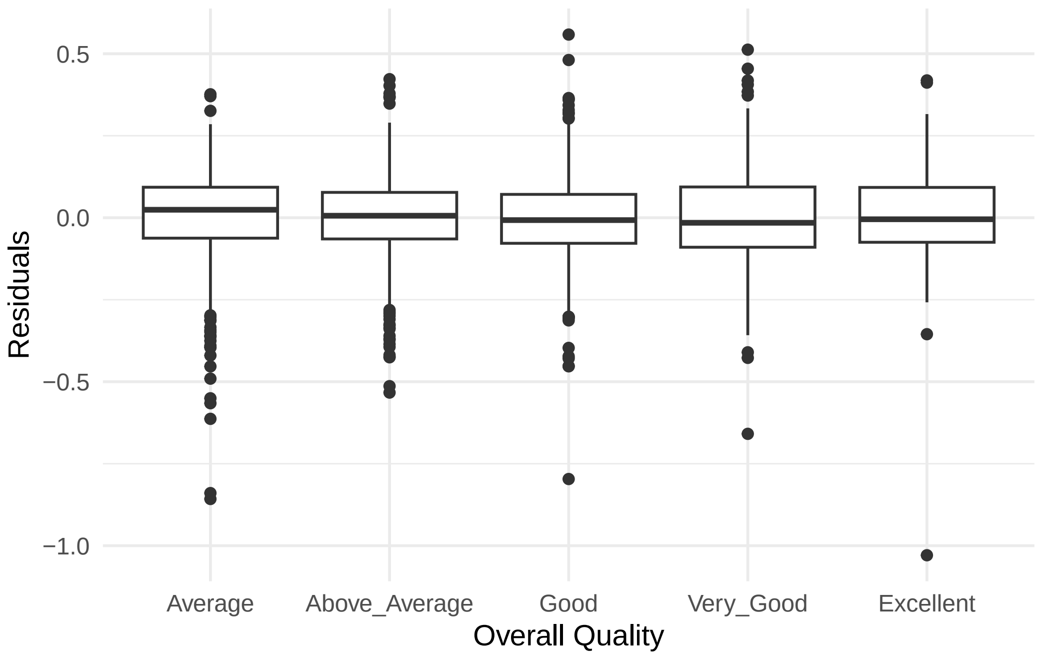
```
`geom_smooth()` using formula = 'y ~ x'
```



```
[[4]]
```


0.5
0.0
−0.5
−1.0
Residuals
2
3
4
Bedrooms Above Ground

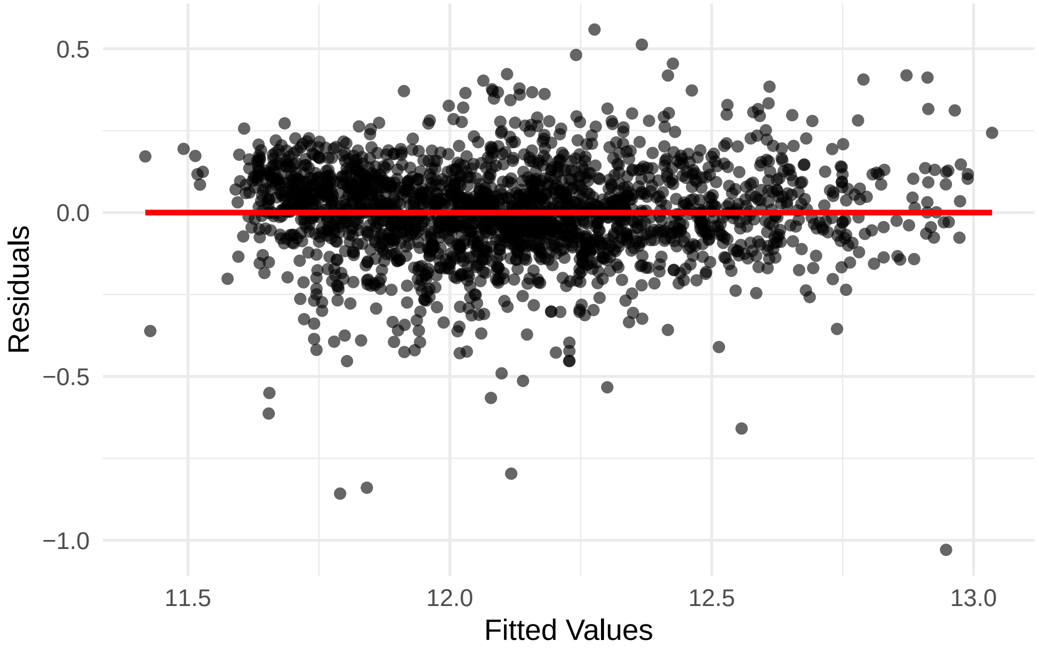
```
[[5]]
```


0.5
0.0
−0.5
−1.0
Residuals
Average
Above_Average
Good
Very_Good
Excellent
Overall Quality

5 Model Fit and Validation

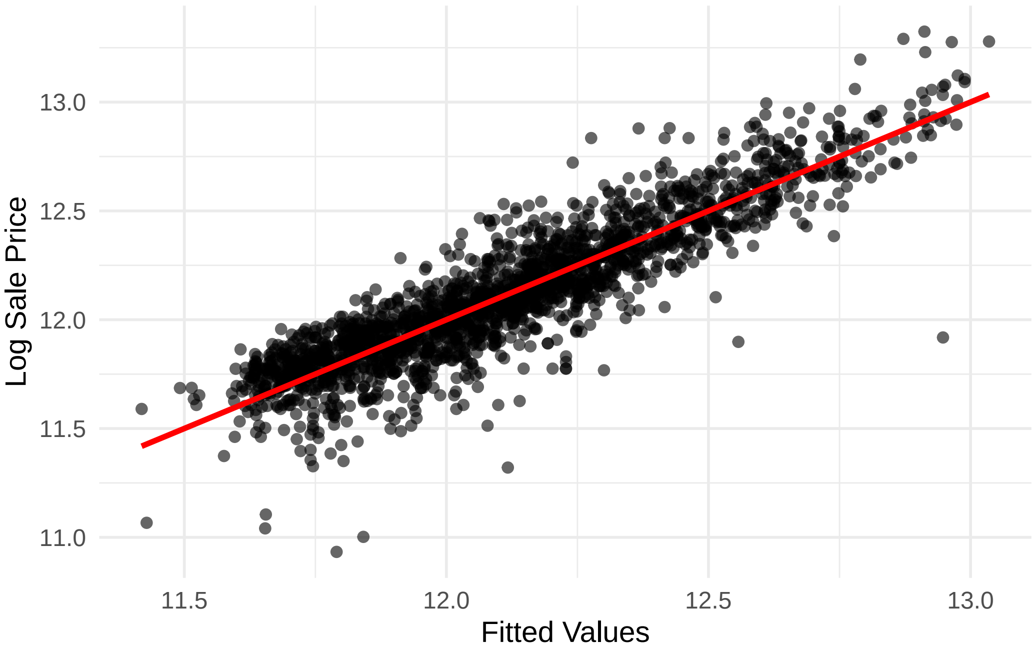
```
ames_data$fitted_values <- fitted(lm_model)

# Fitted vs residuals
ggplot(ames_data, aes(x = fitted_values, y = residuals)) +
  geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Fitted Values", y = "Residuals") + theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



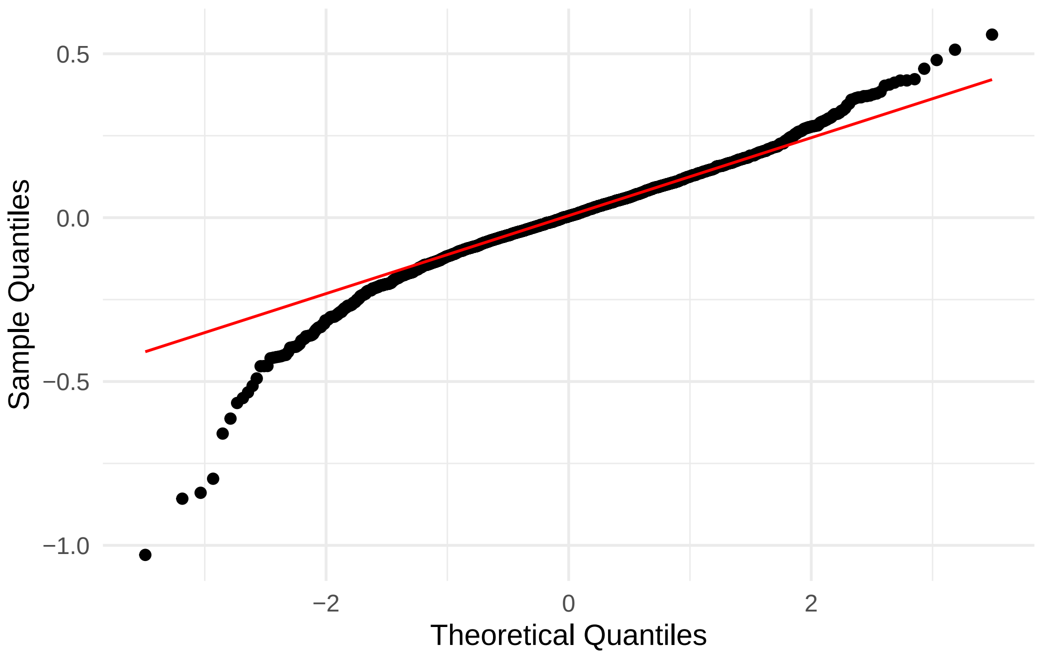
```
# Fitted vs actual
ggplot(ames_data, aes(x = fitted_values, y = log_Sale_Price)) +
  geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Fitted Values", y = "Log Sale Price") + theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



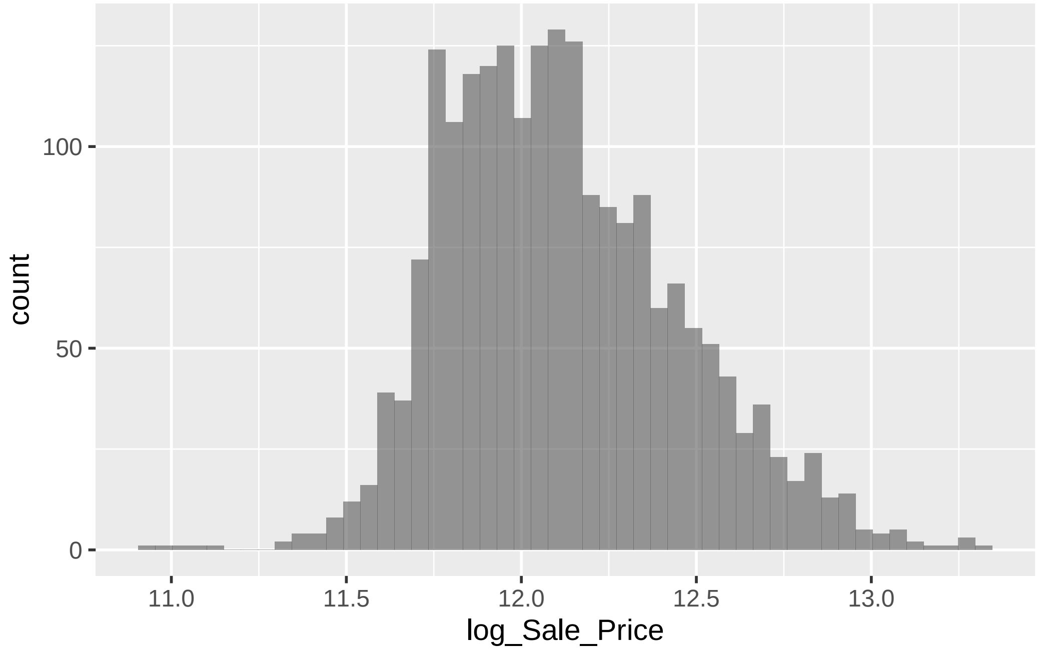
6 QQ Plot

```
ggplot(data = data.frame(residuals = resid(lm_model)), aes(sample = residuals)) +
  stat_qq() + stat_qq_line(color = "red") +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles") + theme_minimal()
```

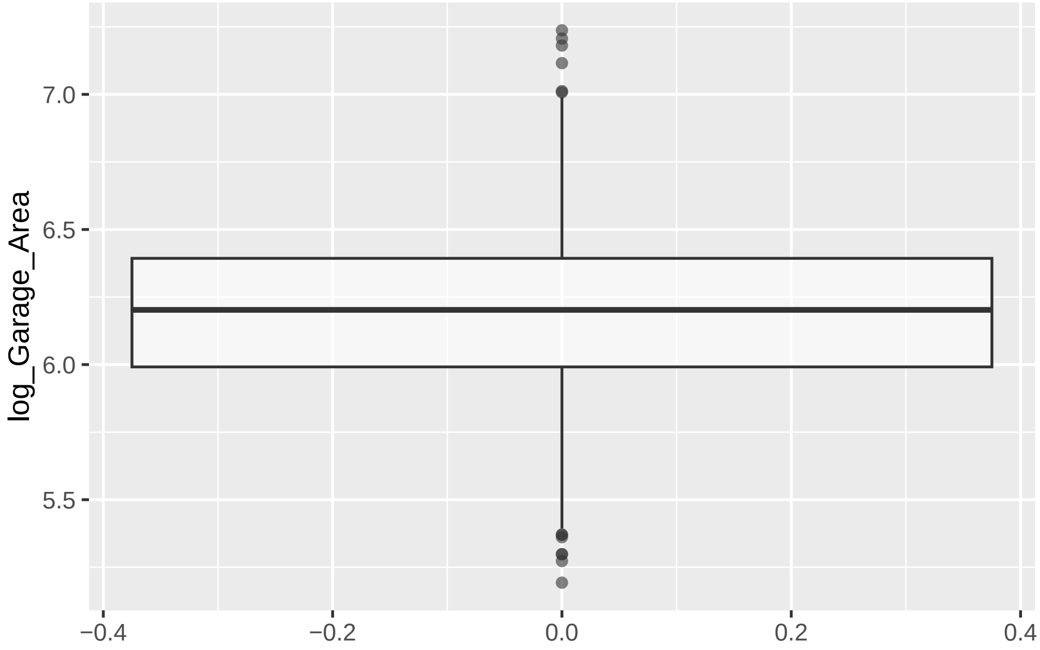


7 Distribution Checks

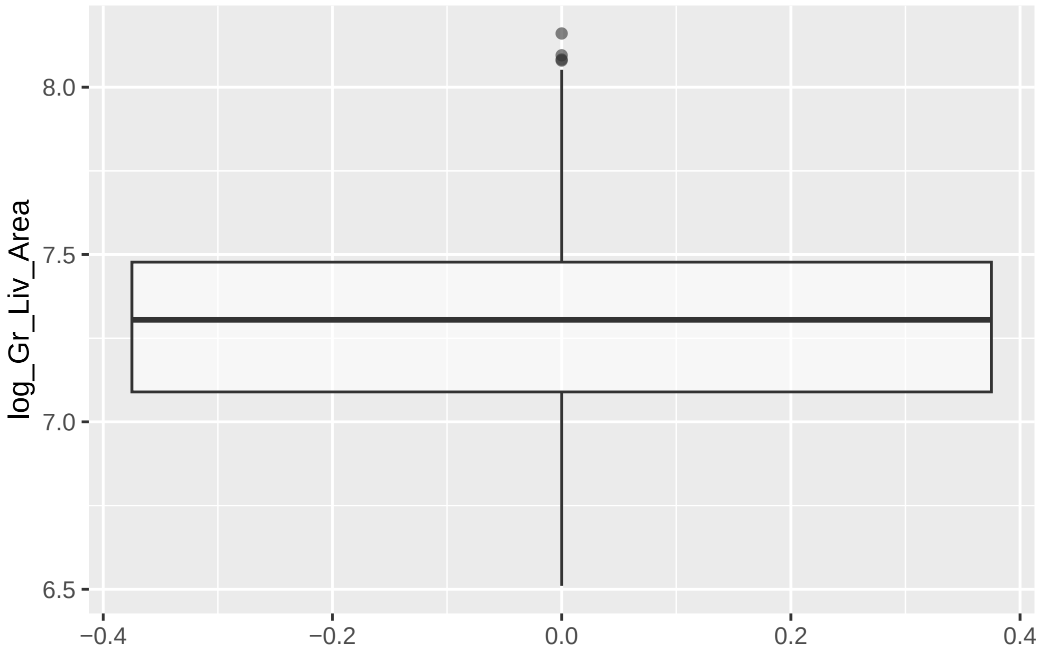
```
ggplot(ames_data, aes(x = log_Sale_Price)) + geom_histogram(bins = 50, alpha = 0.6)
```



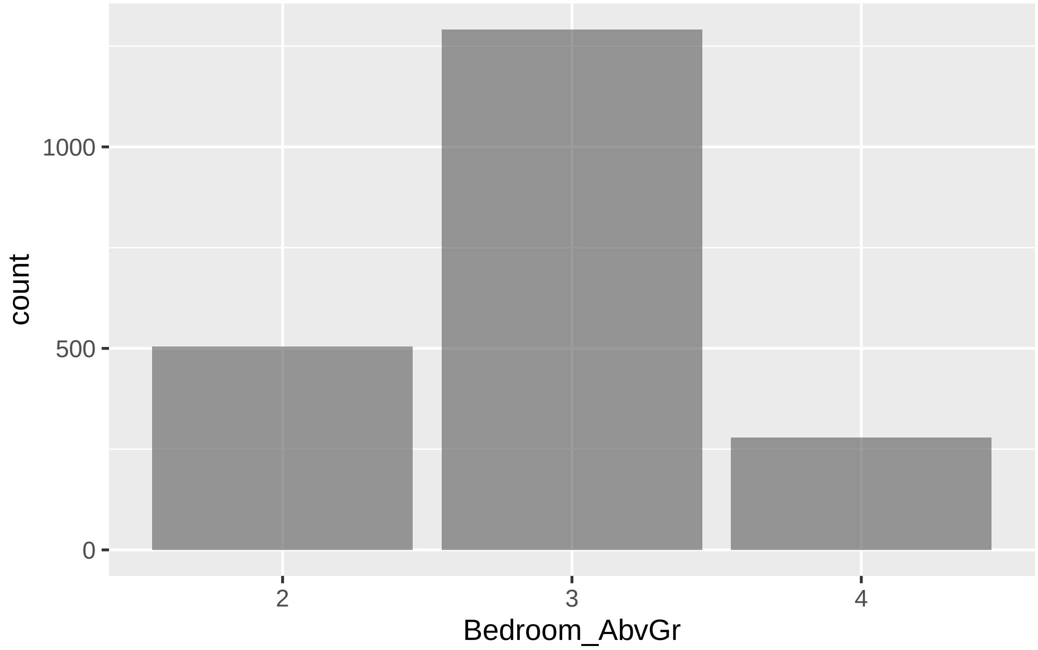
```r
ggplot(ames_data, aes(y = log_Garage_Area)) + geom_boxplot(alpha = 0.6)
```



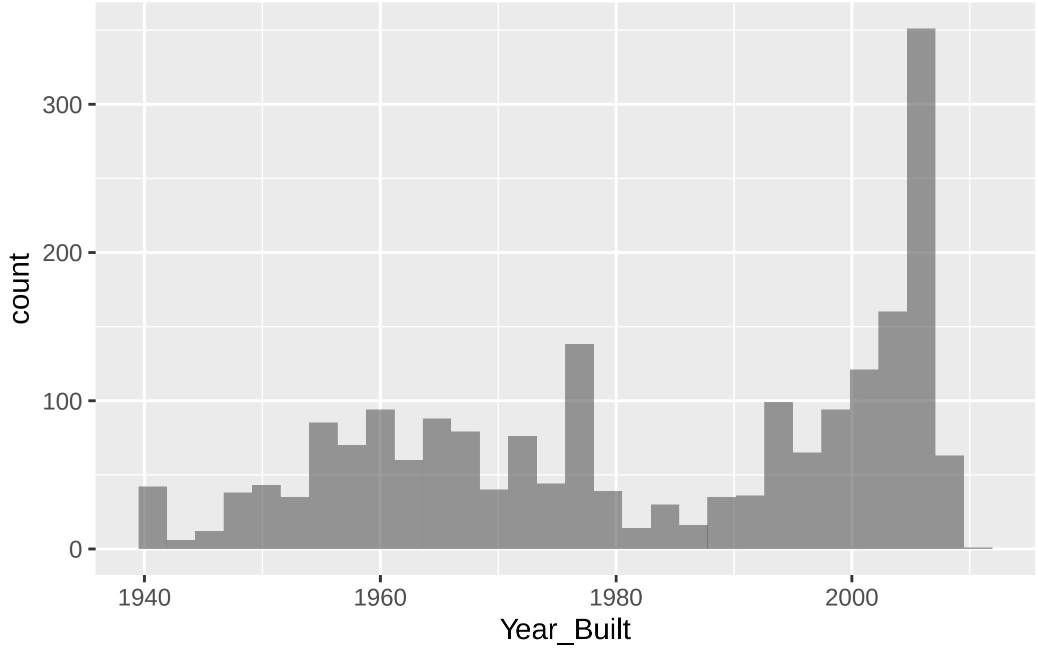
```
ggplot(ames_data, aes(y = log_Gr_Liv_Area)) + geom_boxplot(alpha = 0.6)
```



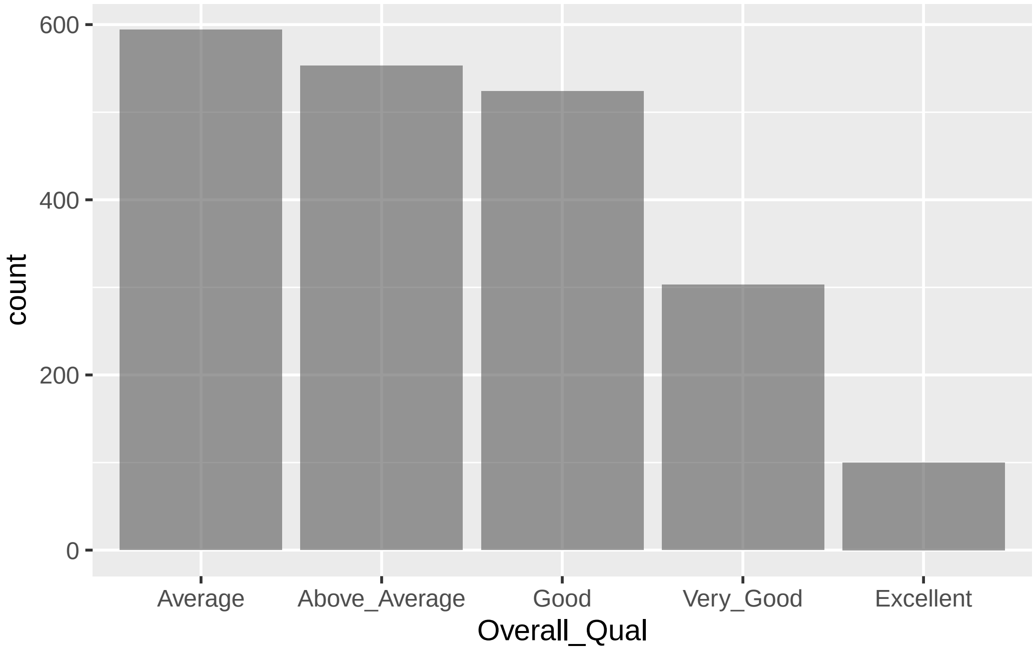
```
ggplot(ames_data, aes(x = Bedroom_AbvGr)) + geom_bar(alpha = 0.6)
```



```
ggplot(ames_data, aes(x = Year_Built)) + geom_histogram(bins = 30, alpha = 0.6)
```



```r
ggplot(ames_data, aes(x = Overall_Qual)) + geom_bar(alpha = 0.6)
```



8 Final Model Summary Table

```r
coefs <- summary(lm_model)$coefficients[, c("Estimate", "Std. Error", "Pr(>|t|)")]
conf_int <- confint(lm_model)
final_model_table <- cbind(
  Estimate = coefs[, "Estimate"],
  Std_Error = coefs[, "Std. Error"],
  CI_Lower = conf_int[, 1],
  CI_Upper = conf_int[, 2],
  p_value = coefs[, "Pr(>|t|)"]
)
round(final_model_table, 4)
```

```
                          Estimate Std_Error CI_Lower CI_Upper p_value
(Intercept)                 2.6813    0.4588   1.7816   3.5810  0.0000
log_Gr_Liv_Area             0.5325    0.0178   0.4976   0.5674  0.0000
log_Garage_Area             0.1480    0.0122   0.1240   0.1719  0.0000
```

```
Year_Built                    0.0023    0.0002   0.0018   0.0028  0.0000
Bedroom_AbvGr3               -0.0156    0.0080  -0.0313   0.0001  0.0515
Bedroom_AbvGr4               -0.0721    0.0134  -0.0984  -0.0459  0.0000
Overall_QualAbove_Average     0.0322    0.0094   0.0138   0.0507  0.0006
Overall_QualGood              0.1131    0.0122   0.0892   0.1370  0.0000
Overall_QualVery_Good         0.2689    0.0151   0.2392   0.2986  0.0000
Overall_QualExcellent         0.4833    0.0204   0.4432   0.5233  0.0000
```

9 Model Performance Metrics

```
r_squared <- summary(lm_model)$r.squared
adj_r_squared <- summary(lm_model)$adj.r.squared
model_aic <- AIC(lm_model)

cat("R-squared:", round(r_squared, 4), "\n")
```

```
R-squared: 0.819
```

```
cat("Adjusted R-squared:", round(adj_r_squared, 4), "\n")
```

```
Adjusted R-squared: 0.8182
```

```
cat("AIC:", round(model_aic, 2), "\n")
```

```
AIC: -2184.5
```

10 Correlation Matrix

correlation_data <- ames_data[, c(“log_Gr_Liv_Area”, “log_Garage_Area”, “Year_Built”)]
cor(round(cor(correlation_data), 3))