Brandon Kokin, Ayrat Aymetov, Saadman Chowdhury

An Empirical Analysis of Housing Prices and Above-Expected Sales in Ames, Iowa

Abstract

Understanding the determinants of housing prices requires separating systematic, theory-driven sources of value from idiosyncratic deviations that arise in individual transactions. Using the Ames Housing dataset, this study develops a regression-based benchmark for expected housing prices and then examines which homes sell above or below that benchmark. Ordinary least squares (OLS) regression on log-transformed sale prices is used to model expected value based on structural characteristics and neighborhood effects. An indicator for above-expected sales is then constructed and analyzed using logistic regression, with model performance evaluated through classification metrics and ROC analysis. Finally, nonparametric bootstrap methods are used to quantify uncertainty in the mean housing price. The results show that quality, size, and location explain a substantial proportion of price variation, while deviations from expected value are more difficult to predict and exhibit considerable noise. These findings highlight both the strengths and limitations of regression-based valuation models in housing markets.

1. Introduction

Housing markets are characterized by substantial heterogeneity. Even within a single city, homes differ widely in physical characteristics, neighborhood context, and perceived quality. While economic theory suggests that many of these differences should be capitalized into sale prices, actual transaction prices often deviate from what would be predicted by observable features alone. These deviations may reflect unobserved characteristics, buyer sentiment, market timing, or bargaining dynamics.

The primary goal of this project is to model housing prices in Ames, Iowa, using a theory-driven set of predictors, and then to examine which homes sell above versus below their model-implied expected price. Rather than focusing solely on prediction accuracy, the analysis emphasizes interpretability and statistical reasoning. A linear regression framework is first used to establish a baseline model of expected prices. Building on this benchmark, a binary outcome indicating above-expected sales is constructed and analyzed using logistic regression and classification diagnostics. In addition to regression modeling, formal analysis of variance (ANOVA) techniques are used to assess the role of neighborhood effects both in isolation and conditional on other structural covariates.

This approach mirrors common practices in applied economics and real estate analytics, where hedonic pricing models are used to estimate expected values and residuals are analyzed to understand unusual or exceptional outcomes. By combining regression modeling, diagnostic analysis, classification methods, ANOVA analysis, and bootstrap inference, the study provides a comprehensive statistical examination of housing price behavior.

2. Data Description

The analysis uses the Ames Housing dataset, which contains 2,930 observations and 82 variables describing residential properties in Ames, Iowa. The dataset includes a rich mix of quantitative variables (such as living area and basement size) and qualitative variables (such as neighborhood and housing style). The response variable of interest is the final sale price of each home.

2.1 Key Variables

Based on economic intuition and prior literature on hedonic pricing models, the following variables are emphasized:

- **Primary sources of variation**:
    - Neighborhood
    - Overall quality of the home
    - Above-ground living area
- **Secondary sources of variation**:
    - Total basement square footage
    - Garage capacity
    - House style

Neighborhood is treated as a high-cardinality categorical variable capturing location-specific pricing. Overall quality is an ordinal measure reflecting construction and finish quality, while living area and basement size measure usable space. Garage capacity proxies for parking availability, and house style captures architectural design.

3. Data Preprocessing and Exploratory Analysis

Prior to modeling, the data underwent several preprocessing steps to ensure consistency with statistical assumptions and economic interpretation. These preprocessing steps ensure that subsequent regression and ANOVA analyses are conducted on variables with appropriate scale, ordering, and interpretation.

3.1 Data Types and Ordinal Variables

Several quality-related variables in the dataset are ordinal in nature. Variables measured on a 1–10 scale (such as Overall Quality and Overall Condition) are retained as numeric, preserving their natural ordering. Other quality variables expressed as letter grades (e.g., Poor to Excellent) are cast as ordered factors to reflect their ordinal structure.

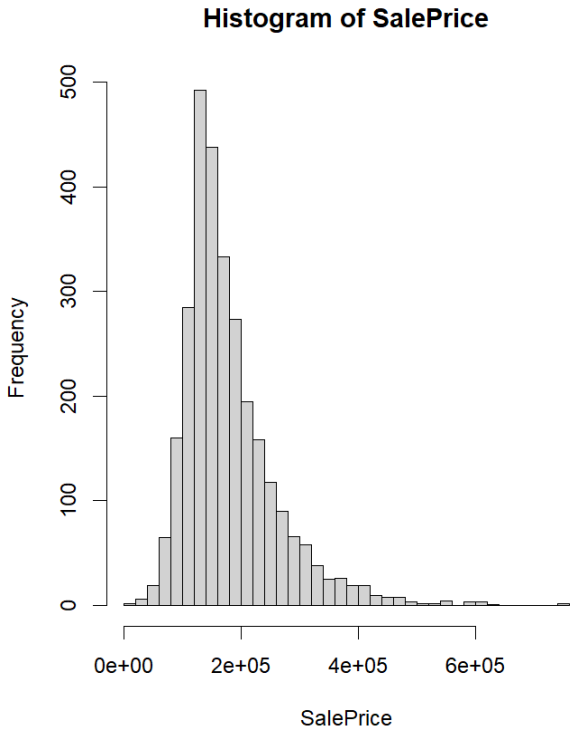
3.2 Handling Missing Values

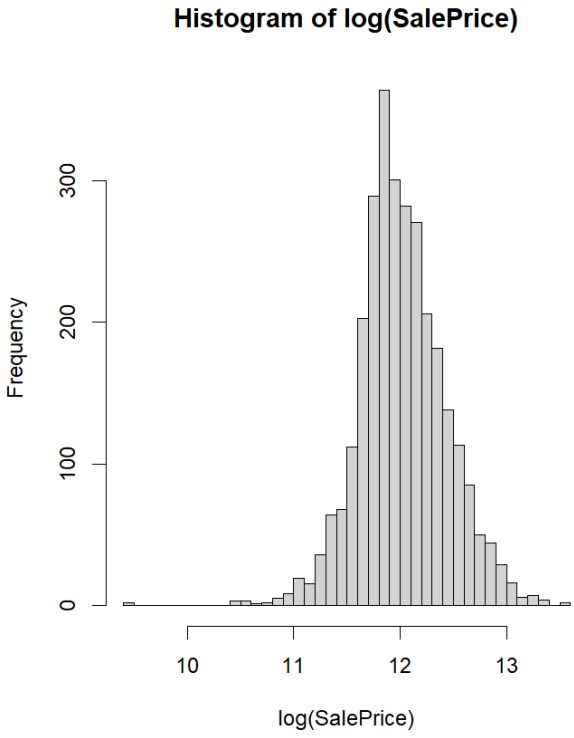
Missing values in housing data can arise for different reasons. In this dataset, missing basement or garage values often indicate the absence of that feature rather than unobserved data. Dropping such observations would remove meaningful information and potentially bias results. Accordingly, missing basement square footage and garage capacity are recoded as zero, indicating no basement or no garage.

Observations with missing sale prices are removed, as the response variable is required for supervised modeling. After this step, no missing values remain in the response.

3.3 Distribution of Sale Prices

An exploratory examination of sale prices reveals substantial right skewness. This is common in housing data, where a small number of high-end properties sell at very high prices. To address this skewness and better satisfy the assumptions of linear regression, the sale price is log-transformed. The log-transformed distribution is much closer to symmetric and supports approximate normality of residuals.





4. OLS Regression for Expected Housing Prices

4.1 Model Specification

The baseline model of expected housing prices is estimated using ordinary least squares regression with log-transformed sale price as the response variable. The model includes neighborhood fixed effects, overall quality, above-ground living area, total basement square footage, garage capacity, and house style.

This specification reflects a standard hedonic pricing framework, where price is modeled as a function of structural characteristics and location. Because the regression includes multiple correlated predictors, subsequent inference relies on Type II ANOVA, which evaluates each variable's marginal contribution conditional on all others.

4.2 Model Results

The OLS model explains a large proportion of the variation in log sale prices, with an adjusted R-squared of approximately 0.84. Overall quality and above-ground living area exhibit the largest t-statistics, confirming their dominant role in determining housing prices. Neighborhood effects remain statistically significant even after controlling for structural characteristics, indicating substantial location-based price premia.

House style variables are generally not statistically significant once size, quality, and neighborhood are accounted for, suggesting that architectural style contributes limited additional explanatory power beyond these core features.

```
Call:
lm(formula = logSalePrice ~ Neighborhood + Overall.Qual + Gr.Liv.Area +
    Total.Bsmt.SF + Garage.Cars + House.Style, data = AmesHousing)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82872 -0.03055  0.00489  0.03882  0.30770

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.679e+00  1.768e-02 264.634  < 2e-16 ***
NeighborhoodBlueste -4.790e-02  2.596e-02  -1.845 0.065077 .
NeighborhoodBrDale  -9.752e-02  1.905e-02  -5.119 3.27e-07 ***
NeighborhoodBrkSide -3.631e-02  1.553e-02  -2.338 0.019469 *
NeighborhoodClearCr  5.899e-02  1.717e-02   3.435 0.000601 ***
NeighborhoodCollgCr  2.940e-02  1.399e-02   2.102 0.035656 *
NeighborhoodCrawfor  5.230e-02  1.519e-02   3.442 0.000585 ***
NeighborhoodEdwards -4.034e-02  1.468e-02  -2.748 0.006036 **
NeighborhoodGilbert  2.337e-02  1.460e-02   1.601 0.109418
NeighborhoodGreens  -1.725e-03  2.811e-02  -0.061 0.951077
NeighborhoodGrnHill  2.232e-01  5.133e-02   4.348 1.42e-05 ***
NeighborhoodIDOTRR  -1.036e-01  1.579e-02  -6.559 6.41e-11 ***
NeighborhoodLandmrk -4.287e-02  7.127e-02  -0.601 0.547577
NeighborhoodMeadowV -7.948e-02  1.843e-02  -4.313 1.66e-05 ***
NeighborhoodMitchel  1.003e-02  1.505e-02   0.667 0.505008
NeighborhoodNAmes   -5.017e-03  1.396e-02  -0.359 0.719341
NeighborhoodNoRidge  5.739e-02  1.602e-02   3.582 0.000346 ***
NeighborhoodNPkVill -4.866e-02  1.985e-02  -2.452 0.014276 *
NeighborhoodNridgHt  6.804e-02  1.444e-02   4.714 2.55e-06 ***
NeighborhoodNWAmes   1.103e-03  1.474e-02   0.075 0.940381
NeighborhoodOldTown -7.542e-02  1.456e-02  -5.179 2.39e-07 ***
NeighborhoodSawyer  -2.650e-03  1.478e-02  -0.179 0.857695
NeighborhoodSawyerW  4.965e-03  1.481e-02   0.335 0.737457
NeighborhoodSomerst  3.787e-02  1.430e-02   2.648 0.008149 **
NeighborhoodStoneBr  6.614e-02  1.658e-02   3.989 6.80e-05 ***
NeighborhoodSWISU   -4.803e-02  1.743e-02  -2.755 0.005906 **
NeighborhoodTimber   4.898e-02  1.563e-02   3.135 0.001737 **
NeighborhoodVeenker  4.406e-02  1.956e-02   2.253 0.024322 *
Overall.Qual         4.677e-02  1.612e-03  29.020  < 2e-16 ***
Gr.Liv.Area          1.014e-04  4.809e-06  21.080  < 2e-16 ***
Total.Bsmt.SF        4.897e-05  4.895e-06  10.003  < 2e-16 ***
Garage.Cars          2.853e-02  2.390e-03  11.939  < 2e-16 ***
House.Style1.5Unf   -2.631e-02  1.688e-02  -1.559 0.119121
House.Style1Story    3.656e-03  5.405e-03   0.677 0.498771
House.Style2.5Fin   -1.390e-02  2.615e-02  -0.532 0.594903
House.Style2.5Unf   -1.749e-02  1.512e-02  -1.157 0.247344
House.Style2Story   -1.602e-03  5.469e-03  -0.293 0.769650
House.StyleSFoyer    3.391e-02  9.499e-03   3.570 0.000362 ***
House.StyleSLvl      1.645e-02  7.906e-03   2.080 0.037568 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06991 on 2891 degrees of freedom
Multiple R-squared:  0.846,     Adjusted R-squared:  0.844
F-statistic: 418.1 on 38 and 2891 DF,  p-value: < 2.2e-16
```

4.3 Interpretation

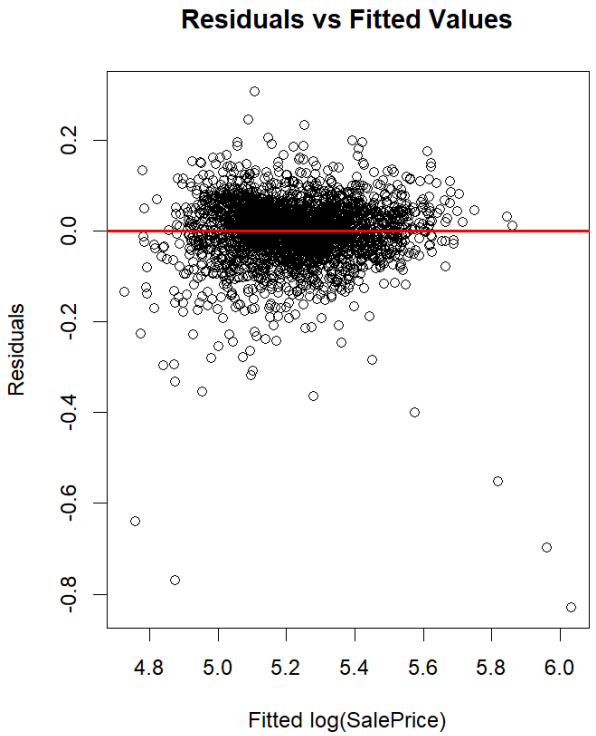
The strong performance of the OLS model supports its use as a benchmark for expected housing prices. Rather than serving purely as a predictive tool, the model provides an economically interpretable decomposition of price variation into systematic components.

5. Model Diagnostics

Regression diagnostics are used to assess whether the assumptions underlying the OLS model are reasonably satisfied.

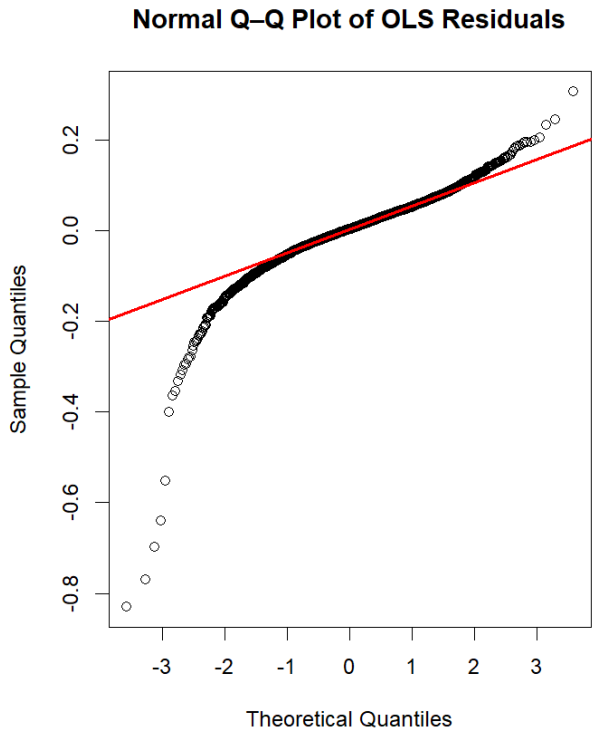
5.1 Residuals vs. Fitted Values

A plot of residuals versus fitted values shows no strong curvature or systematic pattern. Residuals are centered around zero with roughly constant variance, suggesting that the log transformation has effectively stabilized heteroscedasticity.



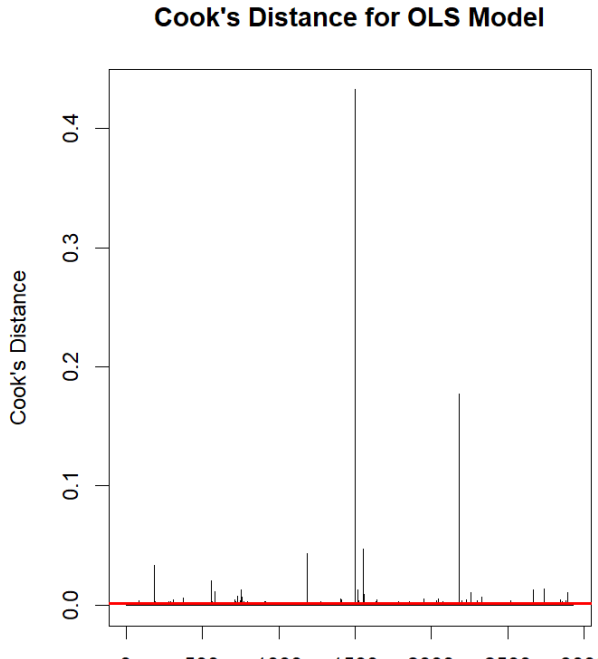
5.2 Normal Q–Q Plot

The normal Q–Q plot indicates that residuals closely follow the reference line in the central region, with modest deviations in the tails. Given the large sample size, these tail deviations are not concerning and do not materially affect inference.



5.3 Influence Diagnostics

Cook’s distance is used to identify potentially influential observations. A small number of observations exceed the conventional threshold, but these appear as isolated cases rather than a systematic issue. No single observation dominates the regression results.



6. Analysis of Variance (ANOVA)

6.1 Type II ANOVA for the Model

The Type II ANOVA results confirm that Overall Quality is the single most important predictor of log-transformed sale prices, followed closely by neighborhood effects. Size-related variables also contribute significantly, while house style explains a smaller but statistically detectable portion of variation when considered jointly.

```
Anova Table (Type II tests)

Response: logSalePrice
              Sum Sq   Df  F value    Pr(>F)
Neighborhood   3.3584   27  25.4507 < 2.2e-16 ***
Overall.Qual   4.1159    1 842.1629 < 2.2e-16 ***
Gr.Liv.Area    2.1718    1 444.3864 < 2.2e-16 ***
Total.Bsmt.SF  0.4891    1 100.0667 < 2.2e-16 ***
Garage.Cars    0.6966    1 142.5382 < 2.2e-16 ***
House.Style    0.1200    7   3.5086 0.0009386 ***
Residuals     14.1290 2891
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6.2 One-Way ANOVA for Neighborhood

A one-way ANOVA testing for differences in mean log(SalePrice) across neighborhoods yields an extremely large F-statistic, providing overwhelming evidence that average housing prices vary substantially by location.

```
              Df Sum Sq Mean Sq F value Pr(>F)
Neighborhood   27  53.46  1.9798   149.9 <2e-16 ***
Residuals    2902  38.32  0.0132
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6.3 Post-Hoc Comparisons (Tukey HSD)

Post-hoc Tukey comparisons reveal that neighborhood-level price differences are widespread rather than isolated. Several neighborhoods exhibit significantly higher mean prices relative to Blmngtn, while others are consistently lower.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = logSalePrice ~ Neighborhood, data = AmesHousing)

$Neighborhood
                         diff           lwr           upr     p adj
Blueste-Blmngtn -0.140127078 -0.2975316466  1.727749e-02 0.1715389
BrDale-Blmngtn  -0.268390632 -0.3806645562 -1.561167e-01 0.0000000
BrkSide-Blmngtn -0.211962270 -0.3025737262 -1.213508e-01 0.0000000
ClearCr-Blmngtn  0.016966034 -0.0863256707  1.202577e-01 1.0000000
CollgCr-Blmngtn  0.001167390 -0.0837078297  8.604261e-02 1.0000000
Crawfor-Blmngtn  0.007412722 -0.0836504070  9.847585e-02 1.0000000
Edwards-Blmngtn -0.196532692 -0.2829102963 -1.101551e-01 0.0000000
Gilbert-Blmngtn -0.014498654 -0.1018284037  7.283110e-02 1.0000000
Greens-Blmngtn  -0.005083857 -0.1763737706  1.662061e-01 1.0000000
GrnHill-Blmngtn  0.150888337 -0.1618428293  4.636195e-01 0.9931965
IDOTRR-Blmngtn  -0.306115909 -0.3982195388 -2.140123e-01 0.0000000
```

7. Constructing the Above-Expected Indicator

Using the fitted values from the OLS model, an indicator variable is constructed that equals one if a home’s observed log sale price exceeds its fitted value and zero otherwise. This variable captures whether a home sells above or below its model-implied expected price.

This construction allows the analysis to shift focus from absolute price levels to relative performance, isolating deviations that may reflect unobserved factors or market dynamics.

8. Logistic Regression for Above-Expected Sales

8.1 Model Specification

A logistic regression model is estimated with the above-expected indicator as the response variable and the same set of predictors used in the OLS model. This methodology was selected to maintain alignment between the models being utilized.

8.2 Results and Interpretation

The logistic regression results indicate that most neighborhood effects are not strongly significant once the expected price benchmark is accounted for. This is consistent with the fact that neighborhood effects are already incorporated into the OLS fitted values. Overall quality has a negative and statistically significant coefficient, indicating that higher-quality homes are less likely to exceed their predicted price. This result is intuitive: quality is already heavily capitalized into the expected value, leaving less room for upside surprises. Basement square footage shows a positive association with above-expected sales, suggesting that basement space may contribute to upside deviations beyond what is captured in the linear model.

```
Call:
glm(formula = AboveExpected ~ Neighborhood + Overall.Qual + Gr.Liv.Area +
    Total.Bsmt.SF + Garage.Cars + House.Style, family = binomial(link = "logit"),
    data = AmesHousing)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.302e-01  5.159e-01  -0.446 0.655441
NeighborhoodBlueste  4.249e-02  7.582e-01   0.056 0.955306
NeighborhoodBrDale   5.619e-01  5.520e-01   1.018 0.308633
NeighborhoodBrkSide  4.668e-01  4.497e-01   1.038 0.299281
NeighborhoodClearCr  8.585e-02  4.971e-01   0.173 0.862887
NeighborhoodCollgCr -9.080e-02  4.048e-01  -0.224 0.822531
NeighborhoodCrawfor  3.166e-01  4.401e-01   0.719 0.471895
NeighborhoodEdwards  4.488e-01  4.259e-01   1.054 0.292008
NeighborhoodGilbert  1.737e-01  4.223e-01   0.411 0.680884
NeighborhoodGreens   1.106e+00  8.287e-01   1.335 0.181947
NeighborhoodGrnHill  5.848e-01  1.521e+00   0.384 0.700674
NeighborhoodIDOTRR   8.787e-01  4.617e-01   1.903 0.057030 .
NeighborhoodLandmrk -1.128e+01  1.970e+02  -0.057 0.954349
NeighborhoodMeadowV  6.947e-02  5.320e-01   0.131 0.896112
NeighborhoodMitchel  3.757e-01  4.358e-01   0.862 0.388731
NeighborhoodNAmes    3.555e-01  4.042e-01   0.880 0.379107
NeighborhoodNoRidge -9.982e-02  4.639e-01  -0.215 0.829640
NeighborhoodNPkVill  5.365e-01  5.749e-01   0.933 0.350644
NeighborhoodNridgHt  1.305e-01  4.179e-01   0.312 0.754845
NeighborhoodNWAmes   4.510e-01  4.277e-01   1.054 0.291672
NeighborhoodOldTown  6.266e-01  4.224e-01   1.483 0.137980
NeighborhoodSawyer   5.173e-01  4.284e-01   1.207 0.227281
NeighborhoodSawyerW  1.617e-01  4.283e-01   0.377 0.705820
NeighborhoodSomerst  3.796e-01  4.137e-01   0.918 0.358817
NeighborhoodStoneBr  3.119e-01  4.807e-01   0.649 0.516465
NeighborhoodSWISU   -6.899e-02  5.057e-01  -0.136 0.891496
NeighborhoodTimber   5.502e-03  4.518e-01   0.012 0.990284
NeighborhoodVeenker  5.850e-01  5.758e-01   1.016 0.309709
Overall.Qual        -1.602e-01  4.739e-02  -3.380 0.000724 ***
Gr.Liv.Area          2.169e-04  1.421e-04   1.526 0.127050
Total.Bsmt.SF        6.992e-04  1.491e-04   4.691 2.72e-06 ***
Garage.Cars         -5.369e-02  7.012e-02  -0.766 0.443855
House.Style1.5Unf    1.813e-01  4.943e-01   0.367 0.713748
House.Style1Story   -8.051e-03  1.576e-01  -0.051 0.959252
House.Style2.5Fin   -7.244e-01  7.825e-01  -0.926 0.354547
House.Style2.5Unf    8.526e-02  4.447e-01   0.192 0.847937
House.Style2Story    2.815e-01  1.604e-01   1.755 0.079286 .
House.StyleSFoyer    8.798e-02  2.758e-01   0.319 0.749762
House.StyleSLvl      7.051e-02  2.306e-01   0.306 0.759724
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4045.0  on 2929  degrees of freedom
Residual deviance: 3962.9  on 2891  degrees of freedom
AIC: 4040.9

Number of Fisher Scoring iterations: 10
```

9. Classification Performance and ROC Analysis

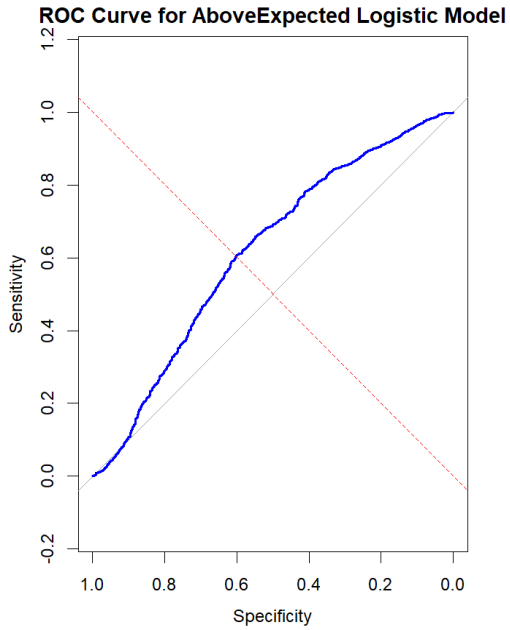
9.1 Confusion Matrix and Metrics

Using a 0.5 probability cutoff, predicted classes are compared to observed outcomes. The resulting confusion matrix yields moderate overall accuracy. Sensitivity is higher than specificity, indicating better performance in identifying homes that sell above expected value than those that do not.

	Actual	
Predicted	0	1
0	550	341
1	804	1235

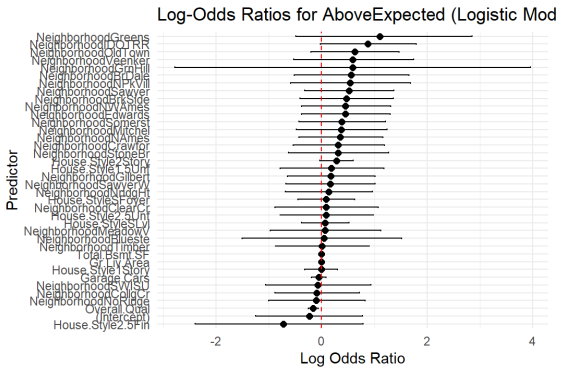
9.2 ROC Curve and AUC

To avoid reliance on an arbitrary cutoff, classifier performance is also evaluated using the receiver operating characteristic (ROC) curve. The area under the curve (AUC) is approximately 0.62, indicating modest but meaningful discrimination beyond random guessing. Given that the outcome variable is constructed from regression residuals and therefore contains substantial noise, this level of performance is reasonable and expected.



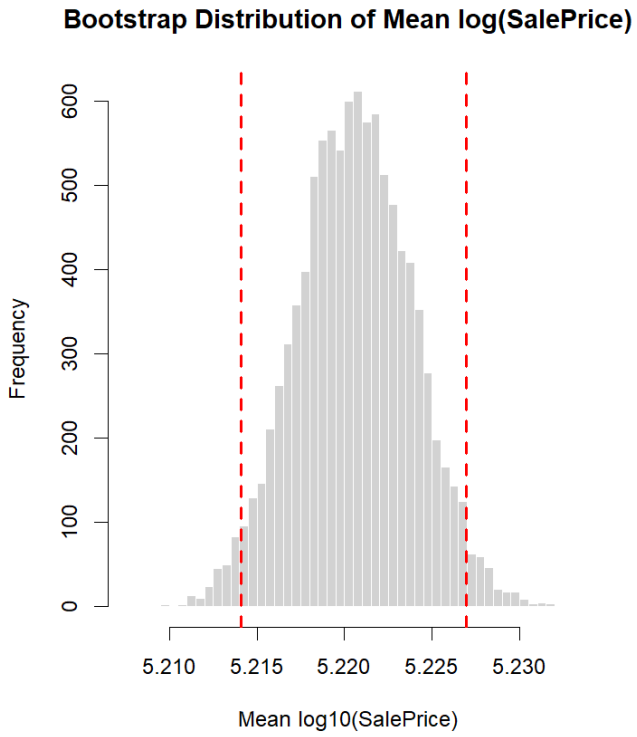
9.3 Log-odds ratio plot

The log-odds ratio plot highlights substantial uncertainty in individual predictors, with most confidence intervals crossing zero. This reinforces the conclusion that above-expected sales are only weakly predictable and largely driven by idiosyncratic factors.



10. Bootstrap Inference for Mean Sale Price

To quantify uncertainty in the mean log-transformed sale price, a nonparametric bootstrap procedure with 10,000 resamples is implemented. The bootstrap distribution of the mean is approximately symmetric and tightly concentrated, indicating stable estimation. A 95% percentile-based confidence interval is constructed on the log scale and then back-transformed to the original price scale. The resulting interval can be interpreted as a confidence interval for the mean sale price.



10. Conclusion

This study applies a range of statistical techniques to analyze housing prices in Ames, Iowa. A theory-driven OLS regression model explains a substantial proportion of price variation and serves as a strong baseline for expected housing values. Diagnostic analysis supports the validity of the linear model assumptions after log transformation. Building on this benchmark, a logistic regression framework is used to examine above-expected sales. While deviations from expected value are inherently noisy, the classification analysis reveals modest but meaningful predictive structure. Finally, bootstrap inference provides a robust assessment of uncertainty in average housing prices. Formal ANOVA and post-hoc inference further demonstrate that neighborhood effects are both statistically and economically meaningful, even before conditioning on structural characteristics.

Overall, the results underscore both the power and limitations of regression-based valuation models. Structural characteristics and location explain most of the systematic variation in housing prices, but residual deviations remain

difficult to predict. This highlights the importance of combining statistical modeling with economic judgment when analyzing real-world housing markets.

11. Appendix

11.1 Project Code

```
if (!requireNamespace("pROC", quietly = TRUE)) {
  install.packages("pROC")
}

if (!requireNamespace("car", quietly = TRUE)) {
  install.packages("car")
}

library(ggplot2)
library(dplyr)
library(pROC)
library(car)

set.seed(123)
```

Module 1: Data Preprocessing

Loading data, cleaning, and preparatory EDA

```{r}
# Loading data, data cleaning, and preparatory Exploratory Data Analysis
# Setup and basic preparation
AmesHousing <- read.csv("data/AmesHousing.csv")

dim(AmesHousing)  # 2930 rows and 82 columns, so very high dimensions make the analysis a bit complex in nature
head(AmesHousing) # head not very useful for types; we'll rely on str() and later cleaning
str(AmesHousing)  # SalePrice is int; OK (R will treat as numeric in modeling)
```

Notice our hypothesized primary and secondary sources of variation based on theory
Primary : Neighborhood, Overall Quality, Above-ground living area
Secondary: Basement Area, Garage Space, Architectural Styl

Note Overall_Qual is actually Ordinal so we elect to change the data type

Note Overall_Cond is actually Ordinal so we elect to change the data type
Note Exter_Cond is actually Ordinal so we elect to change the data type and can also potentially reinput it as a source of variation
Note Bsmt_Cond is actually Ordinal so we elect to change the data type and can also potentially reinput it as a source of variation
we may elect to create a new dataframe with all these to make our analysis smoother, not decided

```{r}
# Quick numeric sanity check on the response
summary(AmesHousing$SalePrice)
```

Theory-driven sources of variation
```{r}
# Primary sources of variation
primary_vars <- c("Neighborhood", "Overall.Qual", "Gr.Liv.Area")

# Secondary sources of variation
secondary_vars <- c("Total.Bsmt.SF", "Garage.Cars", "House.Style")

# Ordinal variables flagged for later recoding
ordinal_vars <- c("Overall.Qual", "Overall.Cond", "Exter.Cond", "Bsmt.Cond")
```

Factor level counts
```{r}
factor_levels <- c()

for (var_name in names(AmesHousing)) {
  x <- AmesHousing[[var_name]]

  # count levels for factor OR character
  if (is.factor(x)) {
    factor_levels[var_name] <- nlevels(x)
  } else if (is.character(x)) {
    factor_levels[var_name] <- length(unique(x))
  }
}

factor_levels <- sort(factor_levels, decreasing = TRUE)
factor_levels
```

```{r}
# Checks
length(AmesHousing$Neighborhood)
length(unique(AmesHousing$Neighborhood))
head(factor_levels)
which.max(factor_levels) # Neighborhood
```
Neighborhood has substantially higher cardinality than other categorical variables, confirming it as a major structural source of variation and motivating careful treatment in regression modeling.

Ordinal recoding (quality variables)
1–10 scales: keep as numeric (already ordinal)
Overall.Qual, Overall.Cond → DO NOTHING (correct as-is)
Quality-grade variables: recode to ordered factors
Ames quality order (worst → best)
```{r}
qual_levels <- c("Po", "Fa", "TA", "Gd", "Ex")

AmesHousing$Exter.Cond <- factor(
AmesHousing$Exter.Cond,
levels = qual_levels,
ordered = TRUE
)

AmesHousing$Bsmt.Cond <- factor(
AmesHousing$Bsmt.Cond,
levels = qual_levels,
ordered = TRUE
)
```

SalePrice distribution checks and transformation
```{r}
hist(AmesHousing$SalePrice,
    breaks = 50,
    main = "Histogram of SalePrice",
    xlab = "SalePrice")                    # clearly this is skewed

# Log-scale version
hist(log(AmesHousing$SalePrice),
    breaks = 50,
    main = "Histogram of log(SalePrice)",
    xlab = "log(SalePrice)")               # much better for normal approximations
```

```
#### Justification for log(SalePrice) transformation
```{r}
# Coefficient of variation comparison
sd(AmesHousing$SalePrice) / mean(AmesHousing$SalePrice)
sd(log(AmesHousing$SalePrice)) / mean(log(AmesHousing$SalePrice))

```
The raw SalePrice distribution is strongly right-skewed, with increasing variance at higher price levels, violating the constant variance assumption of OLS regression. A logarithmic transformation reduces skewness and stabilizes variance, yielding a distribution more appropriate for linear modeling. Accordingly, all subsequent analyses use log(SalePrice) as the response variable.

Preparing data for Analysis procedures
```{r}
AmesHousing <- AmesHousing |> mutate(logSalePrice = log10(SalePrice))
```

Response missingness and structural missingness handling
Observations with missing SalePrice cannot be used in
supervised modeling (OLS, ANOVA, logistic regression).
Such rows are removed if present.
```{r}
AmesHousing <- AmesHousing |>
  filter(!is.na(SalePrice))

sum(is.na(AmesHousing$SalePrice))  # should be 0
```

Handle structural missingness (do NOT drop observations)
For Ames housing data, missing basement or garage values
indicate the absence of that feature, not missing data.
Replacing NA with 0 preserves information and avoids
biased row deletion in downstream OLS analysis.
```{r}
AmesHousing <- AmesHousing |>
mutate(
Total.Bsmt.SF = ifelse(is.na(Total.Bsmt.SF), 0, Total.Bsmt.SF),
Garage.Cars = ifelse(is.na(Garage.Cars), 0, Garage.Cars)
)
```
```

```
sapply(
AmesHousing[c("Total.Bsmt.SF", "Garage.Cars")],
function(x) sum(is.na(x))
)
```

Module 2: OLS Regression — Primary + Secondary Predictors
Model specification
```{r}
AmesHousing$Neighborhood <- factor(AmesHousing$Neighborhood)

ols_model <- lm(
  logSalePrice ~ Neighborhood +
    Overall.Qual +
    Gr.Liv.Area +
    Total.Bsmt.SF +
    Garage.Cars +
    House.Style,
  data = AmesHousing
)

#Model summary
summary(ols_model)
```
The OLS model explains a large proportion of variability in logSalePrice (Adjusted $R^2 \approx 0.84$). Key structural predictors such as Overall.Qual and Gr.Liv.Area are highly statistically significant, supporting the use of this model as a baseline for defining relative over- and under-performance.

Fitted values and residuals
```{r}
AmesHousing$fitted_logPrice <- fitted(ols_model)
AmesHousing$residuals_log   <- resid(ols_model)
```
The OLS results indicate that logSalePrice is strongly explained by structural characteristics and location. Overall.Qual and Gr.Liv.Area exhibit the largest t-statistics, confirming them as dominant drivers of housing prices.

Neighborhood effects remain significant even after controlling for quality and size, indicating location-specific price premia beyond physical attributes.

Several neighborhood indicators are statistically significant, with

both positive and negative effects relative to the reference category (Blmngtn). This suggests meaningful spatial heterogeneity in housing prices that is not fully explained by observable structural features.

Most House.Style categories are not statistically significant once size, quality, and neighborhood are accounted for, indicating that architectural style contributes limited additional explanatory power beyond core structural characteristics.

Reference neighborhood
```{r}
levels(AmesHousing$Neighborhood)
```
Blmngtn is the reference neighborhood in the OLS model, and all neighborhood coefficients are interpreted relative to this baseline.

```{r}
# Multicollinearity check (VIF)
vif(ols_model)
```
The Variance Inflation Factors (GVIF-adjusted for terms with multiple degrees of freedom) all come in below standard cutoffs, so there's no sign of serious multicollinearity issues between the predictors. This means the coefficient estimates from the OLS model should be stable and we can interpret them reliably.

Module 2B: Interaction Check (Neighborhood × Overall.Qual)

```{r}
ols_interaction <- lm(
  logSalePrice ~ Neighborhood * Overall.Qual +
    Gr.Liv.Area +
    Total.Bsmt.SF +
    Garage.Cars +
    House.Style,
  data = AmesHousing
)

summary(ols_interaction)
```

Baseline OLS model retained for defining expected price.
Interaction terms were explored but not adopted due to limited interpretability and model stability concerns.

Module 2C: ANOVA (Type II) for OLS Model
```{r}
Anova(ols_model, type = 2)
```

ANOVA (Type II) Interpretation

The Type II ANOVA assesses the marginal contribution of each predictor after accounting for all other variables in the model.

Neighborhood is highly statistically significant ($p < 2e-16$), confirming strong location-based differences in housing prices.

Overall.Qual has the largest F-statistic, indicating it is the single most important predictor of log(SalePrice).

Gr.Liv.Area, Total.Bsmt.SF, and Garage.Cars are all highly significant, supporting the role of size and functional space in determining prices.

House.Style is statistically significant as a group ($p < 0.001$), although individual style coefficients may be weak, indicating that architectural style contributes modest but non-negligible variation when cons

Module 2D: One-Way ANOVA (Neighborhood Only)

```{r}
anova_neighborhood <- aov(
  logSalePrice ~ Neighborhood,
  data = AmesHousing
)

summary(anova_neighborhood)
```

One-Way ANOVA Interpretation (Neighborhood Only)

The one-way ANOVA tests whether mean log(SalePrice) differs across neighborhoods without controlling for other structural variables.

The F-statistic is very large (F ≈ 150) with p < 2e-16, providing overwhelming evidence that average housing prices differ significantly across neighborhoods.

This confirms Neighborhood as a dominant source of variation in housing prices and motivates its inclusion in the multivariable OLS model.

Module 2E: Post-hoc Comparisons (Tukey HSD)

```{r}
tukey_neighborhood <- TukeyHSD(anova_neighborhood)
tukey_neighborhood
```

A one-way ANOVA followed by Tukey's HSD was used to compare mean log(SalePrice) across all neighborhoods while controlling the family-wise error rate across multiple pairwise comparisons.

The results reveal substantial heterogeneity in housing prices across neighborhoods. Many pairwise differences are statistically significant, indicating that neighborhood-level price effects are widespread rather than isolated.

Relative to Blmngtn (the OLS reference neighborhood), several neighborhoods (e.g., NoRidge, NridgHt, StoneBr) have significantly higher mean log(SalePrice), while others (e.g., BrDale, MeadowV, IDOTRR) have significantly lower mean log(SalePrice).

These post-hoc results complement the OLS findings by illustrating the magnitude and direction of neighborhood price differences in an unconditional setting.

```{r}
boxplot(
  logSalePrice ~ Neighborhood,
  data = AmesHousing,
  las = 2,
  main = "Distribution of log(SalePrice) by Neighborhood",
  ylab = "log10(SalePrice)"
)
```

```

*Boxplot Interpretation*

The boxplot reveals substantial differences in the distribution of log(SalePrice) across neighborhoods. Median prices vary widely, with neighborhoods such as NoRidge, NridgHt, StoneBr, and Somerst exhibiting notably higher typical prices, while BrDale, MeadowV, IDOTRR, and OldTown are among the lowest.

The limited overlap between several neighborhood distributions visually supports the highly significant one-way ANOVA and the Tukey HSD results, indicating that many pairwise neighborhood differences are economically and statistically meaningful.

At the same time, noticeable within-neighborhood variability and the presence of outliers suggest that neighborhood alone does not fully explain housing prices, motivating the inclusion of structural and quality-related covariates in the multivariable OLS model.

## Module 3: Construct AboveExpected indicator

```{r}
AmesHousing$AboveExpected <-
  as.integer(AmesHousing$logSalePrice > AmesHousing$fitted_logPrice)

# Quick check
table(AmesHousing$AboveExpected)
```

Module 4: OLS Diagnostics

Residuals vs Fitted Values

```{r}
plot(
  ols_model$fitted.values,
  ols_model$residuals,
  xlab = "Fitted log(SalePrice)",
  ylab = "Residuals",
  main = "Residuals vs Fitted Values"
)
abline(h = 0, col = "red", lwd = 2)
```
```

```

*Diagnostic check: Residuals vs Fitted Values*

Residuals are densely clustered around zero with no clear curvature,
indicating that the linearity assumption is reasonable.
The spread of residuals appears roughly constant across fitted values,
suggesting no strong evidence of heteroscedasticity after log transformation.
A small number of outlying residuals are present, which is expected in
cross-sectional housing data and does not invalidate the OLS model.

### Module 4B: Normal Q–Q Plot (Normality of Residuals)
```{r}
qqnorm(
  ols_model$residuals,
  main = "Normal Q–Q Plot of OLS Residuals"
)
qqline(ols_model$residuals, col = "red", lwd = 2)
```

Diagnostic check: Normal Q–Q Plot

Residuals follow the reference line closely in the central region,
indicating approximate normality. Deviations occur primarily in the
lower and upper tails, reflecting a small number of extreme observations.
Given the large sample size, these tail deviations are expected and do
not materially violate the normality assumption for OLS inference.

Module 4C: Influence & Leverage (Cook's Distance)
```{r}
plot(
  cooks.distance(ols_model),
  type = "h",
  main = "Cook's Distance for OLS Model",
  ylab = "Cook's Distance",
  xlab = "Observation Index"
)
abline(h = 4 / nrow(AmesHousing), col = "red", lwd = 2)
```

Diagnostic check: Cook's Distance

The Cook's Distance plot shows that most observations have very low
influence on the fitted model. A small number of observations exceed
```

the reference threshold (4/n), indicating potentially influential points. These appear as isolated spikes rather than a widespread pattern, suggesting that no single observation unduly drives the overall regression results.

Module 5: Logistic Regression for AboveExpected

```{r}
logit_model <- glm(
  AboveExpected ~ Neighborhood +
    Overall.Qual +
    Gr.Liv.Area +
    Total.Bsmt.SF +
    Garage.Cars +
    House.Style,
  data = AmesHousing,
  family = binomial(link = "logit")
)

summary(logit_model)
```

Logistic regression Interpretation

The logistic model estimates the probability that a home sells above its OLS-predicted value. Most Neighborhood indicators are not strongly significant once the baseline expected price is accounted for, which is expected given that Neighborhood effects were already absorbed by the OLS fitted values.

Overall.Qual has a statistically significant negative coefficient, indicating that higher-quality homes are less likely to exceed their predicted price, as quality is already strongly incorporated into the expected value benchmark.

Total.Bsmt.SF is positive and statistically significant, suggesting that basement area contributes to upside deviations beyond what is captured by the OLS model.

Other structural variables (Gr.Liv.Area, Garage.Cars, House.Style) show limited additional explanatory power for AboveExpected status.

Module 5B: Predicted Probabilities and Classification
```{r}
# Compute predicted probabilities from the logistic model
AmesHousing$prob_AboveExpected <- predict(
  logit_model,
  type = "response"
)

# Inspect range of predicted probabilities
summary(AmesHousing$prob_AboveExpected)
```
Diagnostic check: Predicted probabilities

The predicted probabilities span a wide range, from values near 0 to values close to 1, indicating that the logistic model provides meaningful discrimination between homes that sell above versus below their expected value. The median and mean probabilities are close to 0.5, which is consistent with the relatively balanced AboveExpected outcome.

Module 5C: Classification Using 0.5 Cutoff
```{r}
# Classify AboveExpected based on predicted probability
AmesHousing$pred_class_0.5 <-
  as.integer(AmesHousing$prob_AboveExpected >= 0.5)

# Confusion matrix
cm <- table(
  Predicted = AmesHousing$pred_class_0.5,
  Actual    = AmesHousing$AboveExpected
)

cm
```

```{r}
# Extract counts programmatically
TN <- cm["0", "0"]
FP <- cm["1", "0"]
FN <- cm["0", "1"]
TP <- cm["1", "1"]

# Compute metrics
accuracy  <- (TP + TN) / sum(cm)
```

```
sensitivity <- TP / (TP + FN)
specificity <- TN / (TN + FP)

accuracy
sensitivity
specificity
```

Diagnostic check: Classification performance at 0.5 cutoff

The classifier achieves moderate overall accuracy.
Sensitivity is relatively high, indicating good ability to
identify homes that sell above their expected value.
Specificity is lower, indicating weaker performance in
identifying homes that do not sell above expected value.
This asymmetry reflects the balanced but noisy nature of
price deviations around the expected benchmark.

Module 5D: ROC Curve and AUC
```{r}
# ROC object using true labels and predicted probabilities
roc_obj <- roc(
  response  = AmesHousing$AboveExpected,
  predictor = AmesHousing$prob_AboveExpected
)

# Plot ROC curve
plot(
  roc_obj,
  main = "ROC Curve for AboveExpected Logistic Model",
  col = "blue",
  lwd = 2
)
abline(a = 0, b = 1, lty = 2, col = "red")

# Compute AUC
auc_val <- auc(roc_obj)
auc_val
```

ROC Curve & AUC Analysis

The ROC curve summarizes classifier performance across all possible
probability cutoffs by plotting sensitivity against 1 − specificity.

This avoids dependence on an arbitrary threshold such as 0.5.

The AUC (Area Under the Curve) measures overall discrimination ability:
the probability that the model assigns a higher predicted probability
to a randomly chosen AboveExpected home than to a BelowExpected home.

An AUC of 0.5 corresponds to random guessing, while values closer to 1
indicate stronger discriminatory power. Here, AUC ≈ 0.62 suggests
modest but meaningful predictive ability beyond chance.

This level of performance is expected given that AboveExpected is defined
as a residual-based outcome and therefore contains substantial noise.

Module 5E: Odds Ratios and Coefficient Visualization
```{r}
# Extract coefficients and confidence intervals
logit_coef <- coef(logit_model)
logit_ci  <- confint(logit_model)

# Convert to odds ratios
odds_ratios <- exp(logit_coef)
odds_ci <- exp(logit_ci)

# Tidy table
or_plot_data <- data.frame(
  Term  = names(odds_ratios),
  OddsRatio = odds_ratios,
  CI_Lower = odds_ci[, 1],
  CI_Upper = odds_ci[, 2],
  row.names = NULL
)

# Remove non-finite values
or_plot_data <- or_plot_data |>
  filter(is.finite(OddsRatio),
         is.finite(CI_Lower),
         is.finite(CI_Upper))

# Log-scale transformation
or_plot_data <- or_plot_data |>
  mutate(
```

```
    logOR = log(OddsRatio),
    logCI_Lower = log(CI_Lower),
    logCI_Upper = log(CI_Upper)
  )

# Plot log-odds ratios with 95% CI
ggplot(or_plot_data,
       aes(x = logOR,
           y = reorder(Term, logOR))) +
  geom_point(size = 2) +
  geom_errorbarh(
    aes(xmin = logCI_Lower, xmax = logCI_Upper),
    height = 0.2
  ) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Log-Odds Ratios for AboveExpected (Logistic Model)",
    x = "Log Odds Ratio",
    y = "Predictor"
  ) +
  theme_minimal()
```

Log-odds ratio interpretation

Points represent estimated log-odds ratios and horizontal lines show 95% confidence intervals from the logistic regression predicting whether a home sells above its OLS-predicted price.

The vertical reference line at 0 indicates no effect on the odds of selling above expected. Most neighborhood coefficients are centered near zero with confidence intervals crossing zero, indicating limited additional neighborhood effects once baseline price expectations are accounted for in the OLS model.

Structural variables such as Total.Bsmt.SF show a small positive effect, while Overall.Qual has a negative log-odds ratio, reflecting that higher quality homes are less likely to exceed their predicted price because quality is already incorporated into the expected value.

Overall, the plot shows substantial uncertainty in individual effects

and confirms that above-expected sales are only weakly predictable,
consistent with the residual-based nature of the outcome.

Module 6: Bootstrap Inference for Mean log(SalePrice)
Using a nonparametric bootstrap to estimate uncertainty
in the mean log-transformed SalePrice. Bootstrapping avoids
reliance on normality assumptions and is appropriate given
the skewed distribution of housing prices.

```{r}
B <- 10000  # number of bootstrap samples
n <- nrow(AmesHousing)

boot_means <- numeric(B)

for (b in 1:B) {
  sample_indices <- sample(1:n, size = n, replace = TRUE)
  boot_sample <- AmesHousing$logSalePrice[sample_indices]
  boot_means[b] <- mean(boot_sample)
}

summary(boot_means)
```

The bootstrap distribution of the mean log(SalePrice)
is approximately symmetric and tightly concentrated,
indicating stable estimation of the population mean.

```{r}
# 95% bootstrap percentile CI on log scale
ci_log <- quantile(boot_means, probs = c(0.025, 0.975))
ci_log

# Back-transform to SalePrice scale
ci_price <- 10^ci_log
ci_price
```

The back-transformed interval provides a 95% confidence
interval for the mean SalePrice on the original dollar scale.
This corresponds to a geometric mean interpretation due
to the log transformation.

```{r}
```

```
#Bootstrap Distribution Visualization
hist(
  boot_means,
  breaks = 40,
  main = "Bootstrap Distribution of Mean log(SalePrice)",
  xlab = "Mean log10(SalePrice)",
  col = "lightgray",
  border = "white"
)
abline(v = ci_log, col = "red", lwd = 2, lty = 2)
```

Illustrative example: interpretation for a single home
```{r example observation}
# Selecting a single example home
example_home <- AmesHousing[1, ]

# Predicting probability of selling above expected price
example_prob <- predict(
logit_model,
newdata = example_home,
type = "response"
)

example_prob
```
For this particular house, the logistic model predicts about a 54% chance that it'll sell for more than what the OLS model expects. So it's slightly more likely than not to beat the predicted price given what we know about the property. This example shows how we can use the model to make predictions for individual homes, not just look at overall patterns.

Summary of Findings
The OLS model captures a substantial portion of the variation in housing prices, with overall quality, above-ground living area, and neighborhood standing out as the key drivers. While neighborhood effects are significant, they don't completely explain price differences once we account for the physical characteristics of the homes.

The logistic regression model predicts whether a home will sell above its expected price and shows moderate discriminatory power (AUC ≈ 0.63). This modest performance makes sense given that we're trying to predict residuals, which are inherently noisy, and it suggests there's limited information beyond what the baseline OLS model already captures.

Bootstrap inference for the mean log(SalePrice) proves to be stable and gives us reliable confidence intervals when we convert back to the dollar scale. Overall, the analysis confirms

that structural features and location are the dominant factors in determining housing prices, while predicting which homes will outperform expectations remains challenging.