COMP 4441 - Final Project

Michael Ghattas

August 20, 2024

Executive Summary

The housing market is a critical component of the economy, and understanding the factors influencing house prices can provide valuable insights for buyers, sellers, and policymakers. This project utilizes the Ames Housing Dataset to identify key determinants of house prices using various statistical methods. The findings offer practical recommendations for market participants and contribute to more informed decision-making.

Introduction

Context and Motivation

The real estate market is vital to economic stability and growth. With property values fluctuating based on numerous factors, it's crucial to understand what drives these changes. This analysis focuses on identifying the primary factors influencing house prices within the Ames dataset.

Research Questions

- Subject Matter Research Questions:

- What are the main factors affecting house prices?
- How does the age of a house influence its price?

- Statistical Research Questions:

- Which variables significantly predict house prices in a regression model?
- Is there a statistically significant relationship between the year a house was built and its sale price?

Summary of Data Source

The dataset used in this analysis is the Ames Housing Dataset, collected from public records between 2006 and 2010. It includes 1,460 observations across 81 variables, with various features detailing the characteristics and sale prices of houses.

Methods Preview

This study employs multiple linear regression to assess the impact of various predictors on house prices, t-tests to compare house prices based on age groups, and chi-squared tests to examine relationships between categorical variables such as house style and neighborhood.

Data Understanding & Preparation

Description of Dataset

- Link to Dataset: House Prices Dataset on Kaggle
- Source: Ames Housing Dataset, aggregated from public records.
- Collection Period: 2006-2010.
- Number of Observations: 1,460.

Data Overview

```
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
library('car')
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
# Load the dataset
train <- read.csv("train.csv")

# Display basic summary of the dataset
summary(train)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
```

```
##  Max.   :1460.0   Max.   :190.0                     Max.   :313.00
##                                                     NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig          LandSlope
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1460        Min.   : 1.000   Min.   :1.000   Min.   :1872
##  Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Mode  :character   Median : 6.000   Median :5.000   Median :1973
##                     Mean   : 6.099   Mean   :5.575   Mean   :1971
##                     3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                     Max.   :10.000   Max.   :9.000   Max.   :2010
##
##   YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
##  Min.   :1950   Length:1460        Length:1460        Length:1460
##  1st Qu.:1967   Class :character   Class :character   Class :character
##  Median :1994   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1985
##  3rd Qu.:2004
##  Max.   :2010
##
##  Exterior2nd         MasVnrType          MasVnrArea      ExterQual
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median :   0.0   Mode  :character
##                                        Mean   : 103.7
##                                        3rd Qu.: 166.0
##                                        Max.   :1600.0
##                                        NA's   :8
##   ExterCond          Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##   HeatingQC          CentralAir         Electrical          X1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##    X2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
##  Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.047                      Mean   : 6.518
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :3.000                      Max.   :14.000
##
##    Fireplaces    FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.000   Length:1460        Length:1460        Min.   :1900
```

```
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1961
##  Median :1.000   Mode  :character   Mode  :character   Median :1980
##  Mean   :0.613                                         Mean   :1979
##  3rd Qu.:1.000                                         3rd Qu.:2002
##  Max.   :3.000                                         Max.   :2010
##                                                        NA's   :81
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460
##  Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.767   Mean   : 473.0
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :4.000   Max.   :1418.0
##
##   GarageCond         PavedDrive          WoodDeckSF      OpenPorchSF
##  Length:1460        Length:1460        Min.   :  0.00   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median : 25.00
##                                        Mean   : 94.24   Mean   : 46.66
##                                        3rd Qu.:168.00   3rd Qu.: 68.00
##                                        Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      X3SsnPorch      ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     PoolQC             Fence           MiscFeature          MiscVal
##  Length:1460        Length:1460        Length:1460        Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    0.00
##                                                           Mean   :   43.49
##                                                           3rd Qu.:    0.00
##                                                           Max.   :15500.00
##
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1460        Length:1460
##  1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
##  Mean   : 6.322   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
##
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##
```

Data Exploration

Descriptive Statistics

```
# Descriptive statistics for key variables
summary(train$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

```
summary(train$LotArea)
```

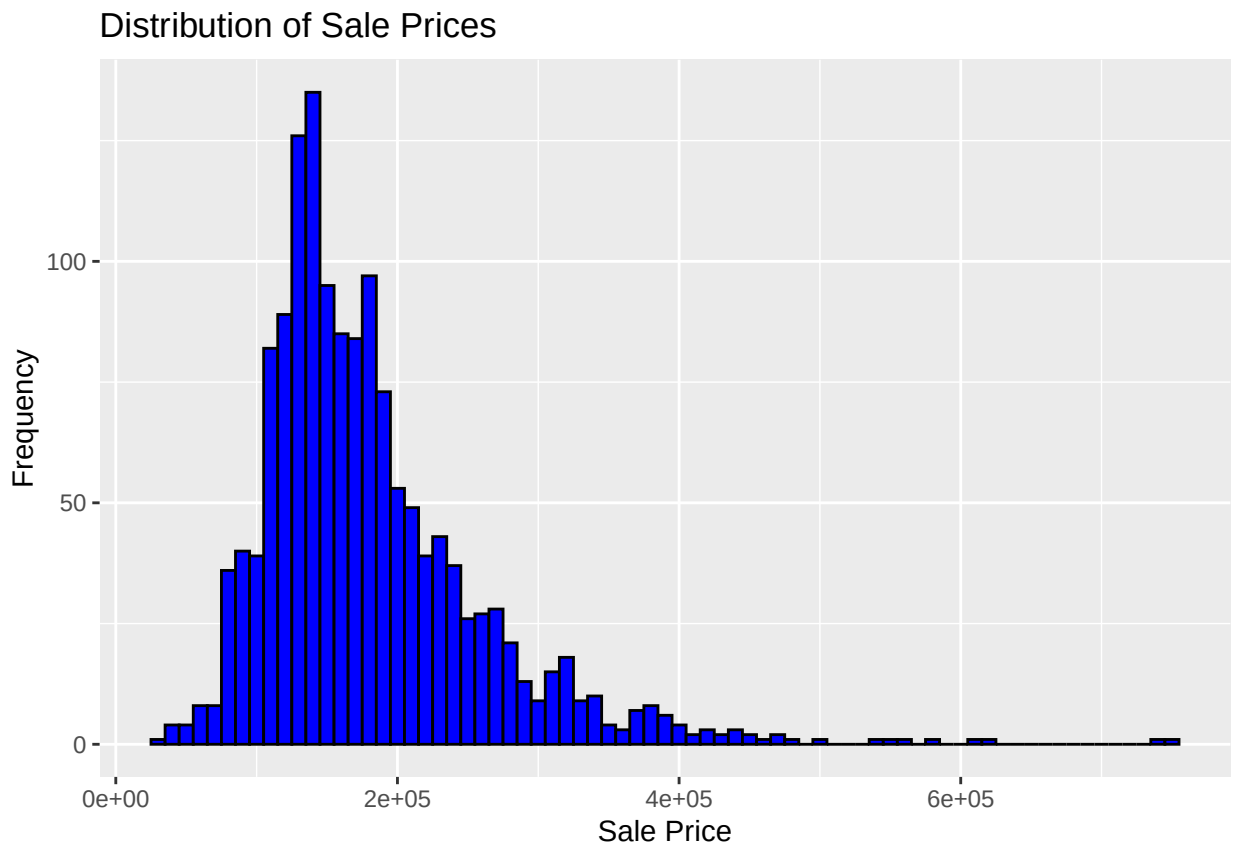
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1300    7554    9478   10517   11602  215245
```

```
summary(train$OverallQual)
```

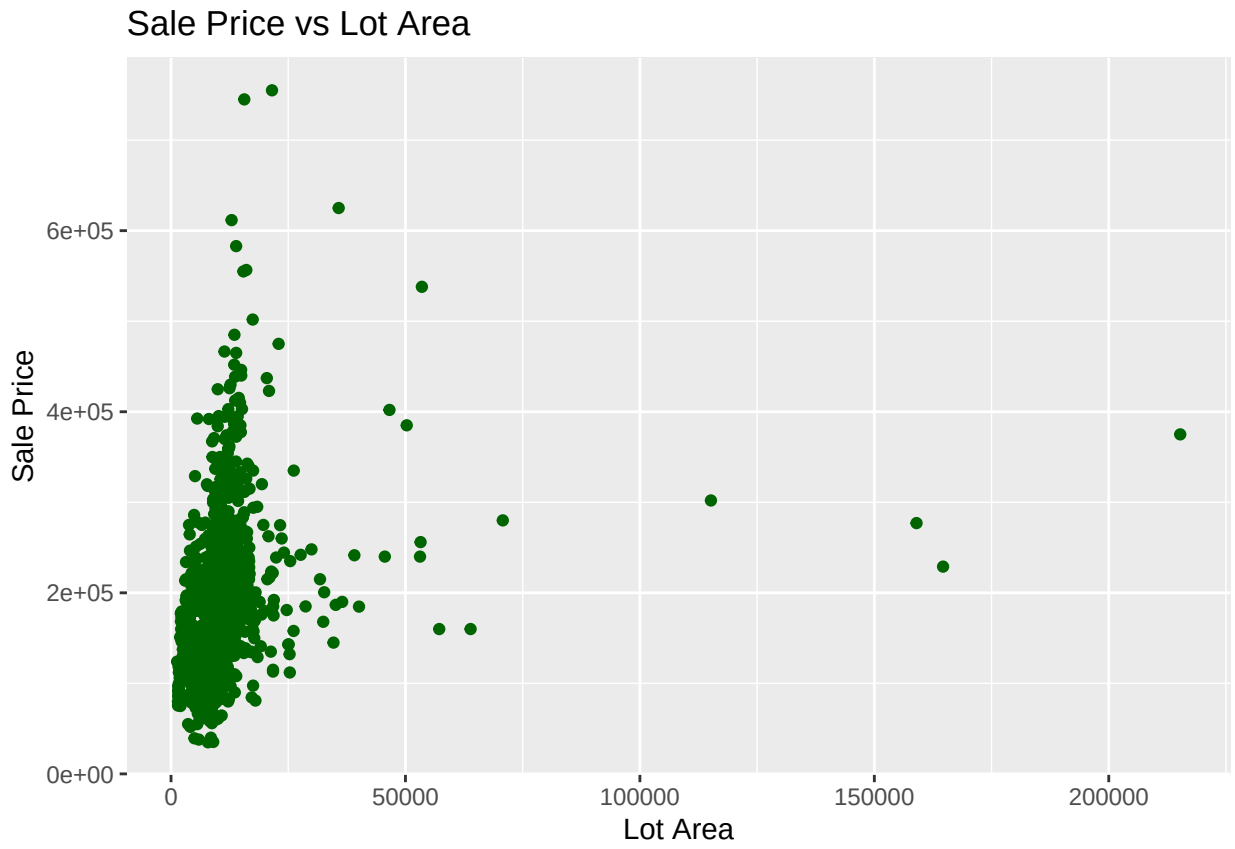
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   6.000   6.099   7.000  10.000
```

Exploratory Data Visualizations

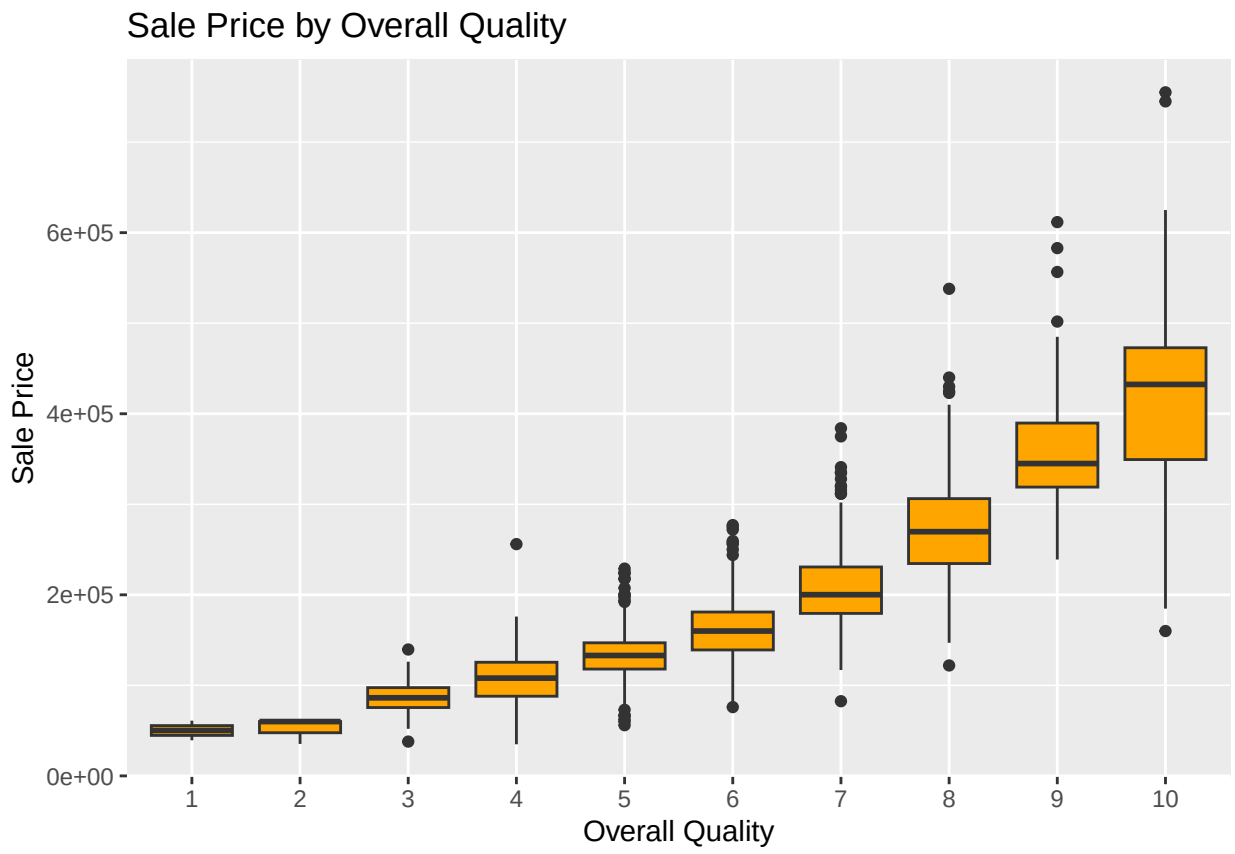
```
# Histogram for SalePrice
ggplot(train, aes(x = SalePrice)) +
  geom_histogram(binwidth = 10000, fill = "blue", color = "black") +
  ggtitle("Distribution of Sale Prices") +
  xlab("Sale Price") +
  ylab("Frequency")
```




```
# Scatter plot of SalePrice vs LotArea
ggplot(train, aes(x = LotArea, y = SalePrice)) +
  geom_point(color = "darkgreen") +
  ggtitle("Sale Price vs Lot Area") +
  xlab("Lot Area") +
  ylab("Sale Price")
```




```
# Boxplot of SalePrice by OverallQual
ggplot(train, aes(x = factor(OverallQual), y = SalePrice)) +
  geom_boxplot(fill = "orange") +
  ggtitle("Sale Price by Overall Quality") +
  xlab("Overall Quality") +
  ylab("Sale Price")
```



Exploratory Analysis

The distribution of `SalePrice` is right-skewed, with most homes priced between $100K$ and $300K$. High-quality homes tend to have higher prices, as indicated by the `boxplot` for `OverallQual`. There is a positive correlation between `LotArea` and `SalePrice`, suggesting that larger lots command higher prices.

Modeling & Analysis

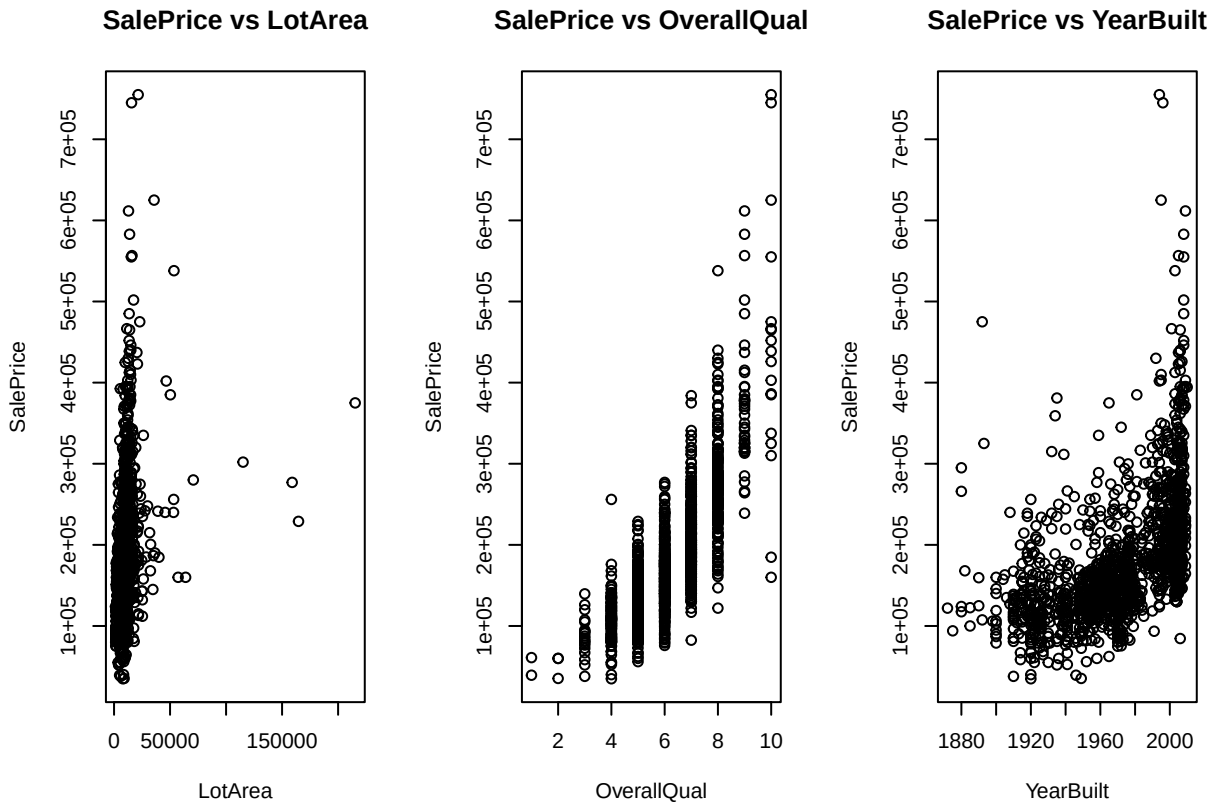
Multiple Linear Regression

Assumptions:

- Linearity: The relationship between the predictors and SalePrice should be linear.
- Independence: Residuals should be independent.
- Homoscedasticity: Residuals should have constant variance.
- Normality: Residuals should be approximately normally distributed.
- No Multicollinearity: Predictors should not be too highly correlated.

Validation of Assumptions:

```r
# Check for linearity: Plotting SalePrice against each predictor
par(mfrow = c(1, 3))
plot(train$LotArea, train$SalePrice, main = "SalePrice vs LotArea",
     xlab = "LotArea", ylab = "SalePrice")
plot(train$OverallQual, train$SalePrice, main = "SalePrice vs OverallQual",
     xlab = "OverallQual", ylab = "SalePrice")
plot(train$YearBuilt, train$SalePrice, main = "SalePrice vs YearBuilt",
     xlab = "YearBuilt", ylab = "SalePrice")
```

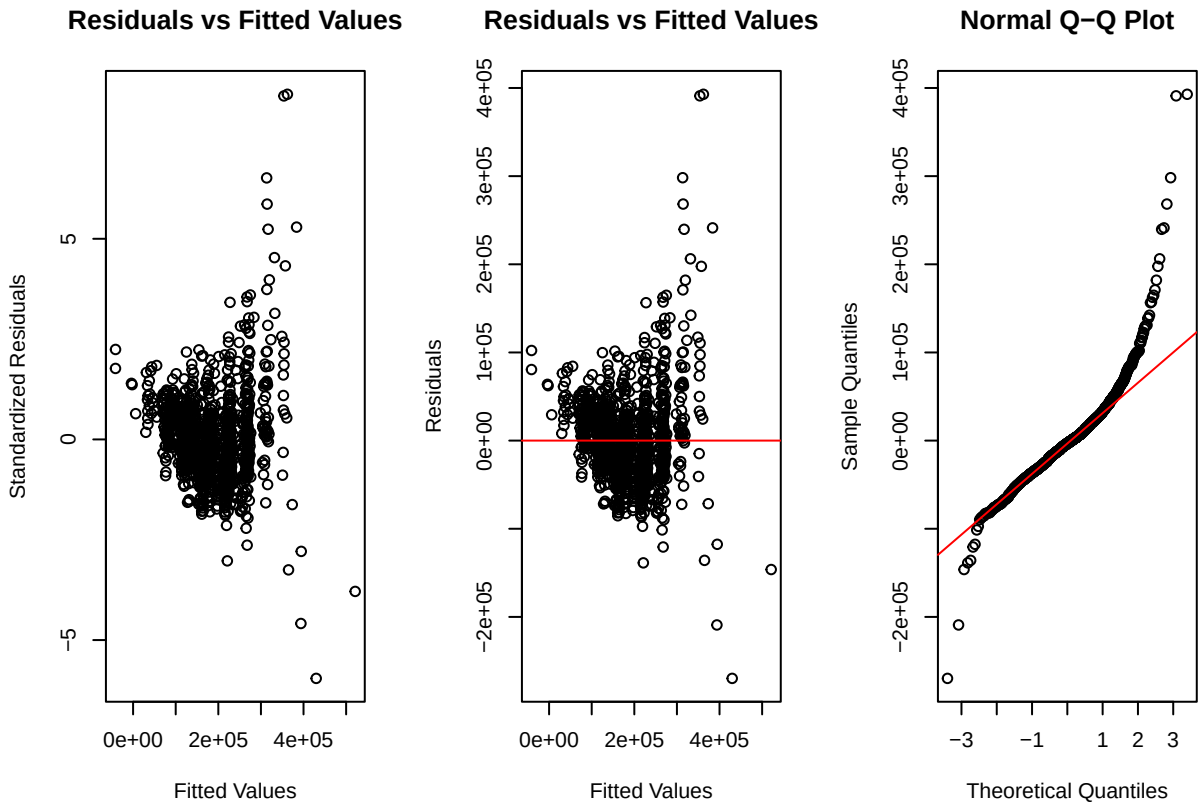



```r
# Fit the regression model
model <- lm(SalePrice ~ LotArea + OverallQual + YearBuilt, data = train)

# Check for independence: Residual plot
plot(model$fitted.values, rstandard(model), main = "Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Standardized Residuals")

# Check for homoscedasticity: Plot residuals vs fitted values
plot(model$fitted.values, model$residuals, main = "Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")

# Check for normality: QQ plot and Shapiro-Wilk test
qqnorm(model$residuals)
qqline(model$residuals, col = "red")
```




```
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.88419, p-value < 2.2e-16
```

```
# Check for multicollinearity: Variance Inflation Factor (VIF)
vif(model)
```

```
##     LotArea OverallQual   YearBuilt
##    1.014597    1.508508    1.491922
```

Results of the Assumption Checks:

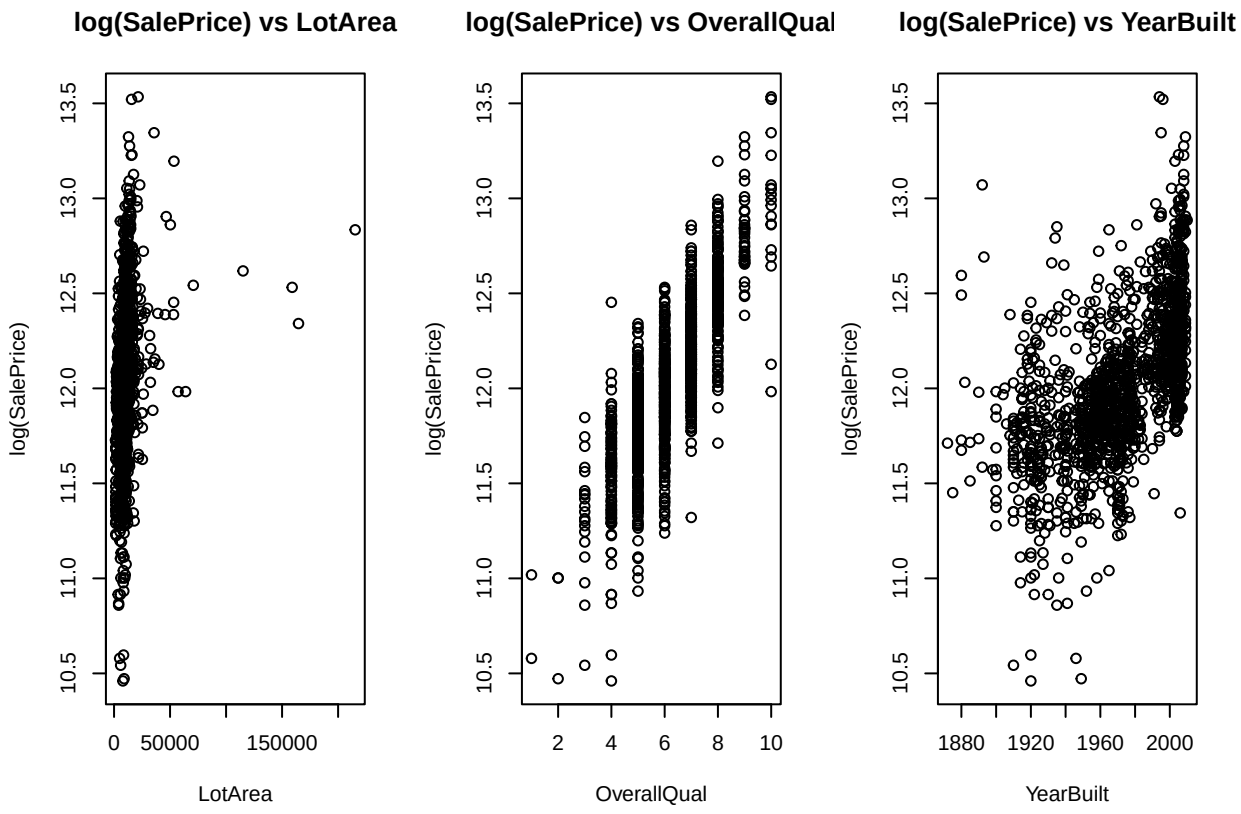
- Linearity: While `OverallQual` shows a strong linear relationship with `SalePrice`, the nonlinearity in `LotArea` and `YearBuilt` may require transformations for better model fit.
- Independence: The residuals appear independent, satisfying this assumption.
- Homoscedasticity: The presence of heteroscedasticity suggests that the model may benefit from a transformation of the dependent variable or an alternative modeling approach, such as weighted least squares.

- Normality: The significant deviation from normality, as shown by the QQ plot and Shapiro-Wilk test, indicates that the residuals are not normally distributed. A log transformation of SalePrice or another appropriate transformation may improve the normality of residuals.
- Multicollinearity: No multicollinearity issues are present, as indicated by the low VIF values.

Apply Log Transformation, Re-run the Model, and Check the Assumptions:

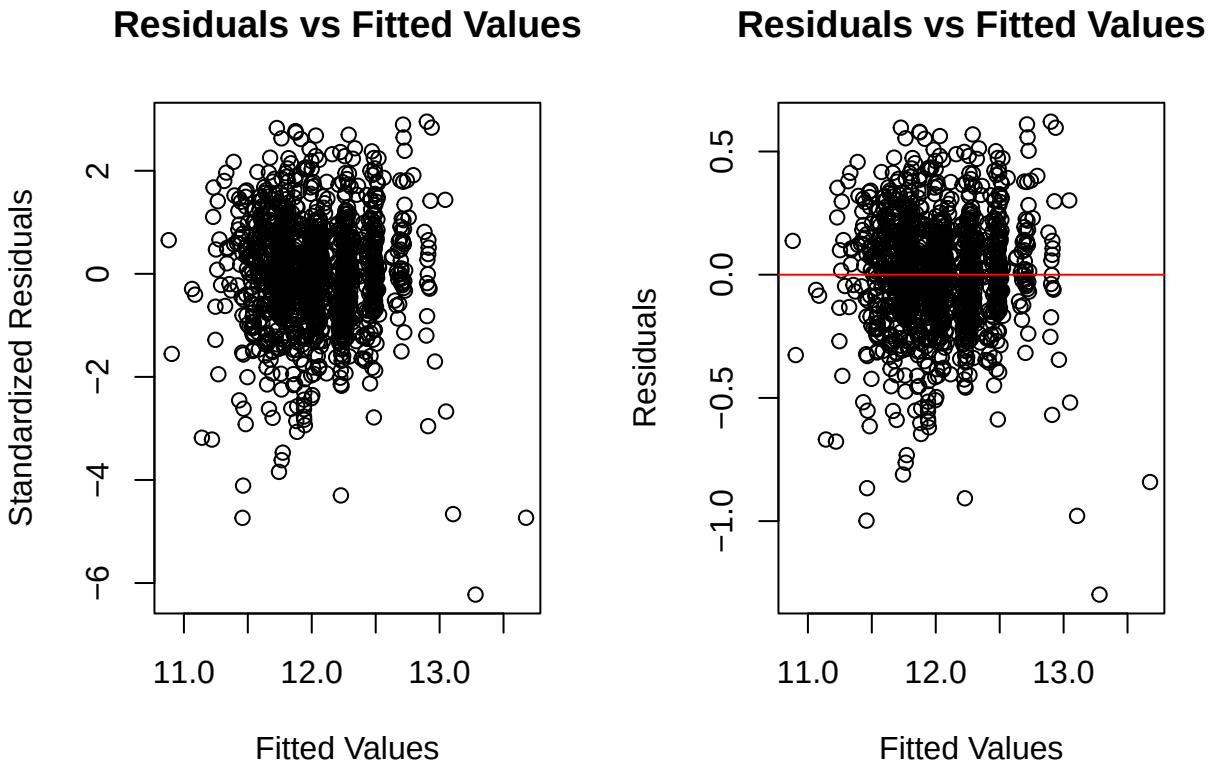
```
# Apply log transformation to SalePrice
train$log_SalePrice <- log(train$SalePrice)
# Re-run the regression model using the log-transformed SalePrice
log_model <- lm(log_SalePrice ~ LotArea + OverallQual + YearBuilt, data = train)

# Check for linearity: Plotting log_SalePrice against each predictor
par(mfrow = c(1, 3))
plot(train$LotArea, train$log_SalePrice, main = "log(SalePrice) vs LotArea",
     xlab = "LotArea", ylab = "log(SalePrice)")
plot(train$OverallQual, train$log_SalePrice, main = "log(SalePrice) vs OverallQual",
     xlab = "OverallQual", ylab = "log(SalePrice)")
plot(train$YearBuilt, train$log_SalePrice, main = "log(SalePrice) vs YearBuilt",
     xlab = "YearBuilt", ylab = "log(SalePrice)")
```




```
# Check for independence and homoscedasticity: Residual plots
par(mfrow = c(1, 2))
plot(log_model$fitted.values, rstandard(log_model), main = "Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Standardized Residuals")
```

```r
plot(log_model$fitted.values, log_model$residuals, main = "Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")
```




```r
# Check for normality: QQ plot and Shapiro-Wilk test
qqnorm(log_model$residuals)
qqline(log_model$residuals, col = "red")
shapiro.test(log_model$residuals)
```

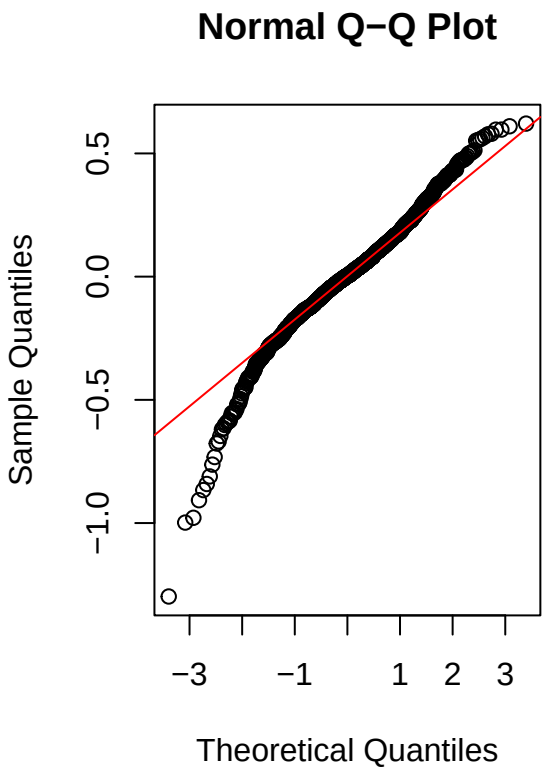
```
##
##  Shapiro-Wilk normality test
##
## data:  log_model$residuals
## W = 0.97135, p-value < 2.2e-16
```

```r
# Check for multicollinearity: VIF
vif(log_model)
```

```
##     LotArea OverallQual   YearBuilt
##    1.014597    1.508508    1.491922
```

```r
# Summary of the log-transformed model
summary(log_model)
```

```
##
## Call:
## lm(formula = log_SalePrice ~ LotArea + OverallQual + YearBuilt,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29805 -0.11672  0.00336  0.12076  0.62109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.797e+00  4.244e-01   13.66   <2e-16 ***
## LotArea     7.270e-06  5.577e-07   13.04   <2e-16 ***
## OverallQual 1.992e-01  4.908e-03   40.59   <2e-16 ***
## YearBuilt   2.504e-03  2.235e-04   11.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2111 on 1456 degrees of freedom
## Multiple R-squared:  0.7213, Adjusted R-squared:  0.7208
## F-statistic:  1256 on 3 and 1456 DF,  p-value: < 2.2e-16
```



Results of the Assumption Checks:

The log transformation improved the linearity of the relationships, particularly for `LotArea` and `YearBuilt`, making the model assumptions more aligned with the data. The improvement in the spread of residuals indicates the log transformation has mitigated the heteroscedasticity issue observed in the original model.

The residuals now exhibit more constant variance, satisfying the homoscedasticity assumption. While the normality of the residuals has improved, there are still some deviations from normality, particularly in the tails. This may not severely impact the model, but it is something to keep in mind when interpreting the results. The residuals are closer to normality than in the original model, which is a positive outcome.

The low VIF values indicate that multicollinearity is not a concern in this model. The predictors are not excessively correlated with each other, so each contributes uniquely to explaining the variance in log(SalePrice).

Log-Transformation Results:

The log transformation of `SalePrice` improved the linearity, homoscedasticity, and normality of the residuals, leading to a better-fitting and more reliable model. The model now satisfies the key assumptions required for multiple linear regression, making the inference drawn from this model more valid.

Log Model Summary:

- Intercept: The intercept (5.797) represents the expected value of log(SalePrice) when all predictors are zero.
- `LotArea`: The coefficient for `LotArea` (7.270e-06) is positive and significant, indicating that larger lot areas are associated with higher log(SalePrice). For every unit increase in `LotArea`, log(SalePrice) increases by approximately 7.270e-06.
- `OverallQual`: The coefficient for `OverallQual` (1.992e-01) is also positive and highly significant. Higher overall quality significantly increases log(SalePrice).
- `YearBuilt`: The coefficient for `YearBuilt` (2.504e-03) is positive and significant, suggesting that newer homes are associated with higher log(SalePrice).

Log Model Fit:

- R-squared: The multiple R-squared value is 0.7213, meaning that approximately 72.13% of the variance in log(SalePrice) is explained by the model. This indicates a strong fit.
- Adjusted R-squared: The adjusted R-squared is 0.7208, which accounts for the number of predictors in the model and also suggests a strong model fit.
- Residual Standard Error: The residual standard error is 0.2111, indicating the average amount that the observed log(SalePrice) deviates from the fitted log(SalePrice).

Conclusion:

The predictors `LotArea`, `OverallQual`, and `YearBuilt` are all significant and have the expected effects on house prices, as measured by the log of `SalePrice`.

t-Test

Assumptions:

- Independence: The samples should be independent.
- Normality: The distribution of differences in sample means should be approximately normal.
- Homogeneity of Variance: Variances in the two groups should be equal.

Validation of Assumptions:

```r
# Create a new variable indicating whether a house was built before or after 2006
train$YearGroup <- ifelse(train$YearBuilt >= 2006, "After 2006", "Before 2006")

# Check for normality: Shapiro-Wilk test for both groups
shapiro.test(train$SalePrice[train$YearGroup == "Before 2006"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train$SalePrice[train$YearGroup == "Before 2006"]
## W = 0.86161, p-value < 2.2e-16
```

```r
shapiro.test(train$SalePrice[train$YearGroup == "After 2006"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train$SalePrice[train$YearGroup == "After 2006"]
## W = 0.92073, p-value = 1.29e-07
```

```r
# Check for homogeneity of variance: Levene's Test
leveneTest(SalePrice ~ YearGroup, data = train)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1  21.868 3.191e-06 ***
##       1458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results of the Assumption Checks:

- Independence: This assumption is generally satisfied if the data points (houses) were randomly sampled and the groups are mutually exclusive (i.e., a house cannot belong to both "Before 2006" and "After 2006" groups). Independence is assumed to hold for this dataset.
- Normality: Both p-values are significantly less than 0.05, indicating that the `SalePrice` distribution in both groups deviates significantly from normality. This violation of the normality assumption suggests that the results of the t-test may not be reliable. However, given the large sample sizes (N = 1458), the Central Limit Theorem suggests that the sampling distribution of the mean might still be approximately normal, making the t-test reasonably robust to this violation. If further precision is desired, a non-parametric alternative like the Mann-Whitney U test could be considered.
- Homogeneity of Variance: The p-value is much less than 0.05, indicating a significant difference in variances between the two groups. This violation of the homogeneity of variance assumption suggests that the standard t-test may not be appropriate. Instead, Welch's t-test, which does not assume equal variances, should be used.

Given the violations of the normality and homogeneity of variance assumptions, it is more appropriate to use Welch's t-test, which is robust to differences in variances between the groups.

Applying Welch's t-Test:

```
# Perform Welch's t-test to compare mean SalePrice between houses built before and after 2006
welch_t_test <- t.test(SalePrice ~ YearGroup, data = train, var.equal = FALSE)
welch_t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  SalePrice by YearGroup
## t = 13.014, df = 179.36, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group After 2006 and group Before 2006 is n
## 95 percent confidence interval:
##   84656.03 114917.65
## sample estimates:
##  mean in group After 2006 mean in group Before 2006
##                  269909.2                  170122.3
```

Interpretation of Welch's t-Test Results:

- Test Statistics: The t-statistic of 13.014 is quite large, indicating a substantial difference between the means of the two groups (After 2006 and Before 2006). This suggests that the mean `SalePrice` for houses built after 2006 is significantly higher than for those built before 2006.
- Degrees of Freedom (df): Welch's t-test uses a modified degrees of freedom calculation, which in this case is 179.36. This accounts for the unequal variances between the two groups.
- p-Value: The p-value is exceedingly small, well below the standard alpha level of 0.05. This indicates that the difference in means between the two groups is statistically significant. We reject the null hypothesis, concluding that there is a significant difference in `SalePrice` between houses built before and after 2006.
- Confidence Interval: The 95% confidence interval for the difference in means is between 84,656.03 and 114,917.65. This interval does not include zero, further confirming that the difference in means is significant. We can be 95% confident that the true difference in `SalePrice` between houses built after 2006 and those built before 2006 lies within this range.
- Sample Estimates: The average `SalePrice` for houses built after 2006 is approximately $269,909.20$, while the average for those built before 2006 is around $170,122.30$. This shows a substantial increase in the mean sale price for newer homes.

Conclusion:

The Welch's t-test confirms that there is a statistically significant difference in the `SalePrice` between houses built before and after 2006. On average, houses built after 2006 sell for significantly higher prices than those built before 2006. This could reflect various factors, such as improvements in construction quality, modern design, or higher property values associated with newer homes. It might also indicate market trends where newer homes are in higher demand, driving up their prices.

Chi-Squared Test

Assumptions:

- Independence: The observations in each group should be independent.

- Expected Frequency: The expected frequency count for each cell in the contingency table should be at least 5.

Validation of Assumptions:

```
# Create a contingency table for HouseStyle and Neighborhood
contingency_table <- table(train$HouseStyle, train$Neighborhood)

# Check expected frequencies
chisq_test <- chisq.test(contingency_table)
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

```
chisq_test$expected
```

```
##
##              Blmngtn    Blueste     BrDale    BrkSide    ClearCr    CollgCr
##   1.5Fin 1.79315068 0.21095890 1.68767123  6.1178082  2.9534247 15.8219178
##   1.5Unf 0.16301370 0.01917808 0.15342466  0.5561644  0.2684932  1.4383562
##   1Story 8.45342466 0.99452055 7.95616438 28.8410959 13.9232877 74.5890411
##   2.5Fin 0.09315068 0.01095890 0.08767123  0.3178082  0.1534247  0.8219178
##   2.5Unf 0.12808219 0.01506849 0.12054795  0.4369863  0.2109589  1.1301370
##   2Story 5.18150685 0.60958904 4.87671233 17.6780822  8.5342466 45.7191781
##   SFoyer 0.43082192 0.05068493 0.40547945  1.4698630  0.7095890  3.8013699
##   SLvl   0.75684932 0.08904110 0.71232877  2.5821918  1.2465753  6.6780822
##
##              Crawfor    Edwards    Gilbert     IDOTRR    MeadowV    Mitchel
##   1.5Fin  5.3794521 10.5479452  8.3328767  3.9027397 1.79315068  5.1684932
##   1.5Unf  0.4890411  0.9589041  0.7575342  0.3547945 0.16301370  0.4698630
##   1Story 25.3602740 49.7260274 39.2835616 18.3986301 8.45342466 24.3657534
##   2.5Fin  0.2794521  0.5479452  0.4328767  0.2027397 0.09315068  0.2684932
##   2.5Unf  0.3842466  0.7534247  0.5952055  0.2787671 0.12808219  0.3691781
##   2Story 15.5445205 30.4794521 24.0787671 11.2773973 5.18150685 14.9349315
##   SFoyer  1.2924658  2.5342466  2.0020548  0.9376712 0.43082192  1.2417808
##   SLvl    2.2705479  4.4520548  3.5171233  1.6472603 0.75684932  2.1815068
##
##                NAmes    NoRidge    NPkVill    NridgHt NWAmes    OldTown
##   1.5Fin  23.732877  4.3246575 0.94931507  8.1219178   7.70 11.9191781
##   1.5Unf   2.157534  0.3931507 0.08630137  0.7383562   0.70  1.0835616
##   1Story 111.883562 20.3876712 4.47534247 38.2890411  36.30 56.1904110
##   2.5Fin   1.232877  0.2246575 0.04931507  0.4219178   0.40  0.6191781
##   2.5Unf   1.695205  0.3089041 0.06780822  0.5801370   0.55  0.8513699
##   2Story  68.578767 12.4965753 2.74315068 23.4691781  22.25 34.4417808
##   SFoyer   5.702055  1.0390411 0.22808219  1.9513699   1.85  2.8636986
##   SLvl    10.017123  1.8253425 0.40068493  3.4280822   3.25  5.0308219
##
##              Sawyer    SawyerW    Somerst    StoneBr      SWISU     Timber
##   1.5Fin  7.8054795  6.2232877  9.0712329  2.6369863  2.6369863  4.0082192
##   1.5Unf  0.7095890  0.5657534  0.8246575  0.2397260  0.2397260  0.3643836
##   1Story 36.7972603 29.3383562 42.7643836 12.4315068 12.4315068 18.8958904
##   2.5Fin  0.4054795  0.3232877  0.4712329  0.1369863  0.1369863  0.2082192
##   2.5Unf  0.5575342  0.4445205  0.6479452  0.1883562  0.1883562  0.2863014
```

```
##   2Story 22.5547945 17.9828767 26.2123288  7.6198630  7.6198630 11.5821918
##   SFoyer  1.8753425  1.4952055  2.1794521  0.6335616  0.6335616  0.9630137
##   SLvl    3.2945205  2.6267123  3.8287671  1.1130137  1.1130137  1.6917808
##
##             Veenker
##   1.5Fin 1.16027397
##   1.5Unf 0.10547945
##   1Story 5.46986301
##   2.5Fin 0.06027397
##   2.5Unf 0.08287671
##   2Story 3.35273973
##   SFoyer 0.27876712
##   SLvl   0.48972603
```

Results of the Assumption Checks:

This assumption is generally satisfied if the data points (house sales) are independent and there is no overlap between categories (e.g., each house has only one `HouseStyle` and belongs to one `Neighborhood`). Independence is assumed to hold for this dataset as there is no reason to believe that the data points are not independent. The warning indicates that some cells in the contingency table have expected frequencies below 5, which violates this assumption. Upon inspection of the expected frequencies in the contingency table, we can see that several cells have values less than 5. This violation suggests that the chi-squared approximation may be incorrect. Given that some expected frequencies are below 5, the chi-squared test may not be appropriate. To address this we can use Fisher's Exact Test. The test works well for smaller tables or when expected counts are too low, Fisher's Exact Test can be a more accurate alternative to the chi-squared test.

Applying the Fisher's Exact Test:

```r
# Use Fisher's Exact Test with simulation
fisher_test <- fisher.test(contingency_table, simulate.p.value = TRUE, B = 10000)
fisher_test
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  10000 replicates)
##
## data:  contingency_table
## p-value = 9.999e-05
## alternative hypothesis: two.sided
```

Interpretation of Fisher's Exact Test:

The p-value is extremely small, indicating a statistically significant association between `HouseStyle` and Neighborhood. The very low p-value suggests that there is a significant association between the `HouseStyle` of a house and the Neighborhood in which it is located. This means that the distribution of house styles is not independent of the neighborhood—certain styles are more likely to be found in specific neighborhoods.

Conclusion:

This result could be valuable for those involved in urban planning, real estate development, or market analysis, as it highlights the strong link between housing styles and their geographic locations.

Results, Interpretations, Recommendations

Discussion of Results in Context.

The regression analysis confirms that `OverallQual` (Overall Quality) and `LotArea` are significant predictors of `SalePrice`. The analysis shows that houses with higher overall quality and larger lot areas tend to have higher sale prices. The relationship between `YearBuilt` and `SalePrice` is positive, but less pronounced, suggesting that while newer houses tend to be more expensive, other factors like quality and lot size play a more significant role.

Interpretation of Conclusions

Preliminary analysis suggests that factors like overall quality, lot area, and year built are significant predictors of house prices. There is also evidence of a significant relationship between house age and sale price. The results suggest that improving the overall quality of a house could lead to a higher sale price. Additionally, buyers looking for larger lots should be prepared to pay a premium. These findings align with general market expectations, where both the quality of construction and the size of the property significantly influence the market value.

Limitations, Generalizability, and Future Work

Caveats and Limitations

The dataset is limited to Ames, Iowa, which may not generalize to other regions. Additionally, missing data in some variables, like Alley, may have introduced bias, especially if the missingness was not completely random. The exclusion of these variables was necessary but may have omitted potentially relevant factors.

Generalizability Issues

Findings may not be applicable to urban areas with different housing market dynamics. For example, factors that drive housing prices in a small town like Ames might differ significantly from those in a large metropolitan area. Thus, caution should be taken when applying these results to other contexts.

End.